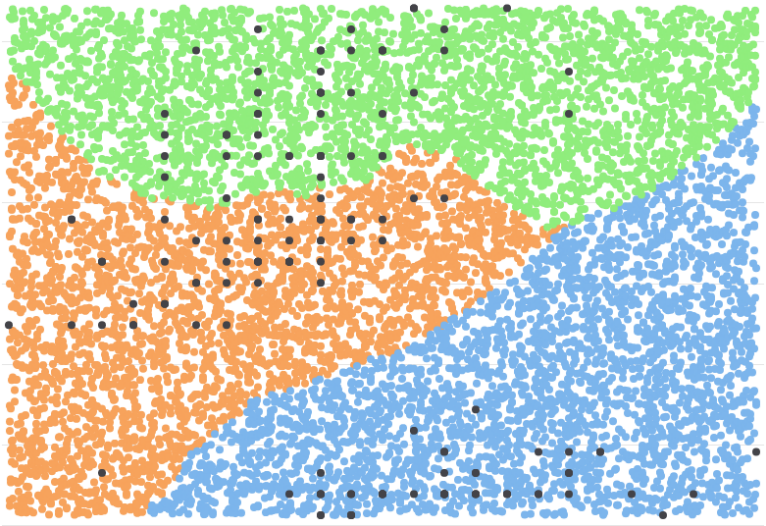
Überwachte Data Mining Verfahren
Klassifikation

Prof. Dr. Hagen Knaf
Studiengang Angewandte Mathematik
Hochschule RheinMain

SS 2023



Visualisierung eines Nächste-Nachbarn-Klassifikators

Hinweis

Vorwort

Im Masterstudiengang »Angewandte Mathematik« werden zwei Lehrveranstaltungen im Fachgebiet »Data Mining« angeboten:

- »Überwachte Data Mining Verfahren«,
- »Unüberwachte Data Mining Verfahren«.

Die Veranstaltungen sind unabhängig voneinander und überlappen sich thematisch nur insoweit, als beide Veranstaltungen allgemeine Ausführungen zum Datenbegriff und zu Data Mining als Arbeitsprozess enthalten.

Das vorliegende Skript zur Veranstaltung »Überwachte Data Mining Verfahren« umfasst die in der Vorlesung behandelten Sachverhalte zum Thema »Klassifikation«. Das methodisch ähnliche Thema »Regression« fehlt in der aktuellen Version.

In Lehrveranstaltungen zum Thema Data Mining spielt die aktiv von allen Teilnehmer:innen durchgeführte Diskussion von Beispielen eine wesentliche Rolle. Nach den Erfahrungen des Dozenten können solche Diskussionen nur in Präsenzveranstaltungen mit der nötigen Intensität geführt werden. Die direkte fachliche Kommunikation Studierender untereinander und mit dem Dozenten unterstützt einerseits das Lernen von Data Mining Methoden und besitzt andererseits erhebliche, berufsvorbereitende Bedeutung, insbesondere für Studierende, die eine Berufstätigkeit im Bereich Datenanalyse anstreben. Branchenunabhängig ist dieser Bereich durch hohe Anforderungen an kommunikative Fähigkeiten und interdisziplinäres Arbeiten geprägt.

Es wird empfohlen zusätzlich zum vorliegenden Skript andere Informationsquellen zum Thema zu nutzen: Das Lehrbuch [FHT], von dem eine legale, freie pdf-Version über die Webseite

https://hastie.su.domains/pub.htm

von einem der Autoren (Trevor Hastie) zur Verfügung gestellt wird, ist ansprechend geschrieben und enthält eine Vielzahl von Beispielen mit frei verfügbaren Daten, sodass diese etwa mit der Software RapidMiner nachvollzogen werden können. Weiter sind die über den YouTube-Kanal

https://www.youtube.com/@TubingenML

der Arbeitsgruppen für Maschinelles Lernen der Universität Tübingen angebotenen Vorlesungen gut als komplementäre Quellen geeignet.

Software: Die im Skript behandelten Beispiele wurden entweder mit der Software *Matlab* der Firma MathWorks oder mit *RapidMiner* bearbeitet. Eine kostenlose Version von RapidMiner kann von der Webseite

https://rapidminer.com/

des Unternehmens Rapidminer heruntergeladen werden.

Anmerkung (Februar 2024): Das Unternehmen Rapidminer ist im Unternehmen Altair aufgegangen. Die Software RapidMiner existiert dort aktuell unter dem Namen *Altair AI Studio* in einer freien Version weiter.

Inhaltsverzeichnis

Einleitung

Nach einer Definition der Wissenschaftler Ian Witten[1] und Eibe Frank bezeichnet man »jede Aktivität, die das Ziel verfolgt, potentiell nützliche Information aus gegebenen Daten zu extrahieren« als Data Mining [WFH]. In diesem allgemeinen Sinn existiert Data Mining bereits mindestens, seitdem es Lebensformen mit Sinnesorganen gibt: Das Auge etwa in Kombination mit einem Komplex von Nervenzellen wie einem Gehirn wird von vielen Lebensformen zur Erkennung von Nahrung benutzt. Die folgenden Ausführungen beschränken sich allerdings auf menschliche Aktivitäten. Auch bei dieser Betrachtungsweise existiert Data Mining bereits seit langer Zeit: Naturforscher wie etwa Carl von Linné (1707 – 1778), Alexander von Humboldt (1769 – 1859) und Charles Darwin (1809 – 1882) sammelten akribisch Daten, um basierend auf deren Analyse Theorien wie beispielsweise die Evolutionstheorie zu entwickeln.

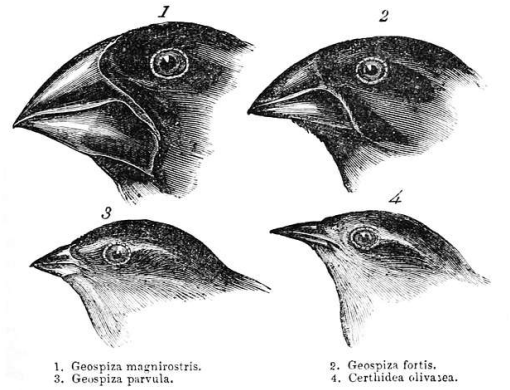


Abbildung 1: Vergleichende Skizzen von Finken der Galapagos-Inseln [Dar]

Der Begriff »Data Mining« wird häufig zusammen mit den Begriffen »Mustererkennung« (»Pattern Recognition«), »Statistisches Lernen« (»Statistical Learning«), »Maschinelles Lernen« (»Machine Learning«) und »Künstliche Intelligenz« (»Artificial Intelligence«) benutzt, wobei wenig auf eine Abgrenzung der Begriffe gegeneinander geachtet wird. Es ist daher möglicherweise nützlich etwas zu diesen Begriffen und ihren wechselseitigen Beziehungen zu sagen.

[1] I. H. Witten, englischer Computerwissenschaftler, 1947 – 2023

Die Analyse von Daten, die entweder »im Feld« oder unter kontrollierten, experimentellen Bedingungen erhoben wurden, ist die Grundlage jeder Naturwissenschaft: Die Ergebnisse von Datenanalysen dienen dabei grob gesprochen der Entwicklung und dem Test von Theorien über bestimmte Bereiche der Realität. Außerhalb des Bereichs der Naturwissenschaften sind Datenanalysen auch für das Funktionieren von Gemeinwesen und deren Interaktionen von essentieller Bedeutung, da ihre Ergebnisse als Richtungsweiser für die Planung und Ausführung von Maßnahmen dienen können. Unabhängig vom Einsatzgebiet dient die Analyse von Daten stets der Entdeckung und quantitativen Beschreibung von Zusammenhängen zwischen Größen des jeweiligen Interesses. Solche Zusammenhänge können deterministisch oder stochastisch sein: Eine oder mehrere Größen legen eine andere entweder eindeutig fest oder erhöhen die Wahrscheinlichkeit dafür, dass letztere Werte in einem bestimmten Bereich annimmt.

Die theoretischen Grundlagen der Datenanalyse werden durch das mathematische Teilgebiet der *Wahrscheinlichkeitstheorie* geliefert, in der der Begriff »Wahrscheinlichkeit« systematisch ausgearbeitet wird unter anderem, um reale Vorgänge mit einer Zufallskomponente quantitativ beschreiben zu können. Letzeres ist für die Analyse von Daten von doppelter Relevanz: Einerseits sind die Sachverhalte oder Prozesse, über die man sich mit Hilfe einer Datenanalyse Informationen verschaffen möchte, häufig nicht deterministisch, sondern besitzen vom Zufall abhängige Komponenten. Andererseits treten bei der Datenerhebung selbst in der Regel zufällige Fehler auf. Konkrete Verfahren zur Analyse gegebener Daten werden durch die *Statistik* geliefert, insbesondere durch das Teilgebiet der *multivariaten Statistik*, in dem es um die Untersuchung des Zusammenhangs zwischen mehreren Zufallsgrößen geht. In der Statistik werden außerdem Verfahren zur Schätzung der Güte von Ergebnissen einer Datenanalyse entwickelt. Sie stützt sich auf die Wahrscheinlichkeitstheorie und ergänzt diese um praktische Verfahren.

Der Begriff »Data Mining« wird in mehrdeutiger Weise verwendet. In der engsten Fassung steht Data Mining für die effiziente, computergestützte Analyse von häufig großen Datenmengen mit statistischen Methoden. Dies schließt die Entwicklung entsprechender Algorithmen ein. Mit steigender Leistungsfähigkeit der Computer wurden und werden auch Verfahren im Data Mining verwendet, die anstelle einer fundierten statistischen eine eher empirische Begründung besitzen. Einige solche Verfahren beruhen beispielsweise auf der »Geometrie der Daten«. An diesem Punkt unterscheidet sich Data Mining von der Datenanalyse mit Verfahren der multivariaten Statistik.

Etwas weiter gefasst schließt Data-Mining auch das Vorbereiten der Daten für die Analyse mit ein. Irreführend an der Begriffswahl ist die Assoziation mit dem Gewinnen von Mineralien oder Kohle durch Bergbau, denn im Data Mining werden *keine* Daten durch »Abbau« gewonnen: Die Daten liegen bereits vor.

Ein Data Mining Prozess verfolgt entweder das Ziel der Bestätigung und quantitativen Beschreibung eines (vermuteten) Zusammenhangs zwischen gegebenen Größen oder der Entdeckung von gänzlich neuen Zusammenhängen zwischen Größen, zu denen jeweils Daten vorliegen. Hier liegt auch die Verbindung zum »Statistisches Lernen«: Nach Vladimir Vapnik [Vap] bezeichnet dieses grob gesprochen das datenbasierte Ermitteln einer Abbildung f innerhalb einer gegebenen Klasse $\mathbf{F}$ von »Modellen«, die einen gegebenen, in der Regel nur stochastischen Zusammenhang zwischen zwei Größen X und Y möglichst gut approximiert. Die Verfahren des statistischen Lernens können also im Data Mining genutzt werden, um Zusammenhänge zwischen Größen näherungsweise zu beschreiben.

»Mustererkennung« bezeichnet die Fähigkeit eines informationsverarbeitenden Systems, Objekte anhand (eines Teils) ihrer Merkmale in Klassen einzuteilen. Das System kann dabei beispielsweise ein Organismus oder ein Computerprogramm sein. Im ersten Fall befasst sich die Kognitionswissenschaft mit der Untersuchung dieser Fähigkeit. Die Entwicklung computergestützter Verfahren zur Mustererkennung dagegen ist in der Schnittmenge zwischen Informatik und Mathematik verortet. Die Verfahren selbst finden direkte Anwendung im Data Mining: Entsprechend der zwei bereits formulierten, allgemeinen Ziele des Data Mining besteht das Klassifikationsproblem darin, datenbasiert eine quantitative Beschreibung einer (vermuteten) Klasseneinteilung von Objekten anhand der Ausprägung bestimmter Merkmale zu ermitteln. Die Beschreibung kann dann zur Klassifikation von Objekten genutzt werden, zu denen keine Klasseninformation vorliegt. Offensichtlich ist dieser Teil der Mustererkennung ein Spezialfall des Szenarios, das im statischen Lernen betrachtet wird: Die Größe X besteht aus der Zusammenfassung der betrachteten Merkmale eines zu klassifizierenden Objekts, während Y die Klasse dieses Objekts angibt. Andererseits wird in dem als *Clusteranalyse* bezeichneten Teilgebiet von Data Mining versucht, datenbasiert Klassen von Objekten zu entdecken. Die dort zum Einsatz kommenden Verfahren stammen ebenfalls aus dem Bereich der Mustererkennung.

Der Begriff »maschinelles Lernen« bezeichnet alle Verfahren, die aus gegebenen Daten eine Abbildung mit bestimmten Optimalitätseigenschaften

ermitteln. Insbesondere umfasst der Begriff das statistische Lernen und die Mustererkennung im Sinn der Informatik, aber auch weitere zum Beispiel auf geometrischen Methoden basierende Verfahren. Die geforderten Optimalitätseigenschaften beziehen sich dabei auf das Verhalten der ermittelten Abbildung bei Anwendung auf neue Daten, die so genannte *Generalisierungsfähigkeit*: Hat man zum Beispiel eine Abbildung anhand gegebener Daten ermittelt, die Objekten aufgrund der Ausprägung bestimmter Merkmale eine von endlich vielen Klassen zuordnet, so soll diese Abbildung auch auf anderen als den gegebenen Daten möglichst häufig eine korrekte Klassenzuweisung bewirken.

Das Fachgebiet der »künstlichen Intelligenz« ist aktuell nur sehr unscharf definiert, was zum Teil daran liegt, dass es keine allgemein anerkannte Definition des Begriffs »Intelligenz« gibt. Das Lexikon der Neurowissenschaften von Spektrum[Spe] definiert: *Die künstliche Intelligenz ist ein Teilgebiet der Informatik, welches sich mit der Erforschung von Mechanismen des intelligenten menschlichen Verhaltens befaßt (Intelligenz). Dieses geschieht durch Simulation mit Hilfe künstlicher Artefakte, gewöhnlich mit Computerprogrammen auf einer Rechenmaschine (Computersimulation).* Historisch waren die Gebiete »künstliche Intelligenz« und »maschinelles Lernen« ursprünglich identisch. Die Fokussierung auf das »Lernen aus Daten« in letzterem hat zu einer Abgrenzung der Gebiete geführt, sodass das Gebiet der »künstlichen Intelligenz« heute das umfassendere ist.

Nachdem das Fachgebiet Data Mining nun in einen größeren Kontext eingeordnet wurde, fehlen noch einige Worte zur Einordnung des Inhalts der Veranstaltung: Im maschinellen Lernen unterscheidet man unter anderem zwischen »überwachtem Lernen« und »unüberwachtem Lernen«. Im ersten Fall wird, wie bereits erläutert, datenbasiert der Zusammenhang zwischen einer Größe X und einer Größe Y durch Wahl einer geeigneten Abbildung f aus einem Modellraum approximiert. Diese Wahl wird als »Lernprozess« betrachtet, der durch Vergleich der Werte von $f(X)$ und Y »überwacht« wird. Dagegen kann eine Lernprozess, der datenbasiert Zusammenhänge erst entdecken soll, nicht durch ein solches Verfahren überwacht werden. In der Veranstaltung »Überwachte Data Mining Verfahren« geht es entsprechend um die Anwendung derjenigen Verfahren aus dem maschinellen Lernen, die in den Bereich der überwachten Lernverfahren gehören.

1 Daten

Der Begriff »Daten« wird in der Umgangssprache in der Regel unpräzise und mit wechselnden Bedeutungen benutzt. Im Gegensatz dazu ist der im Data Mining verwendete Datenbegriff von mathematischer Präzision. Um einen guten Einstieg in die folgende Darstellung zu bekommen, mache sich der Leser Folgendes klar: Daten werden nicht um ihrer selbst willen analysiert. Vielmehr werden mit Daten die Eigenschaften von bestimmten Objekten erfasst, über die man durch die Datenanalyse etwas in Erfahrung bringen möchte. Beispielsweise kann man als Objekte eine Auswahl von Pilzarten betrachten, über die man ein Pilzbestimmungsbuch schreiben möchte. Um die einzelnen Arten zu beschreiben und gegeneinander abzugrenzen erfasst man von möglichst vielen realen Pilzen bekannter Artzugehörigkeit Eigenschaften wie Hutfarbe, Konsistenz der Huthaut, Geruch, Lamellen- oder Röhrenpilz, Auftreten von Verfärbungen bei Druck usw.

Es sei Ω eine Menge von Objekten derselben Art; sie wird häufig als *Population* oder *Grundgesamtheit* bezeichnet. Jedes Objekt $\omega \in \Omega$ werde durch die *Ausprägungen* von $m \in \mathbb{N}$ *Merkmalen* $M_1, \ldots, M_m$ beschrieben. Jedes Merkmal kann mathematisch als Abbildung

$$M_k : \Omega \to S_k$$

aufgefasst werden, die jedem Objekt ω die Ausprägung $M_k(\omega)$ dieses Merkmals zuordnet. Der Wertebereich S_k umfasst also die möglichen Ausprägungen von M_k. Abhängig vom »Typ« der Menge S_k unterteilt man Merkmale grob in drei Kategorien:

- Numerische Merkmale: $S_k \subseteq \mathbb{R}$.

  Beispiel: Die Wuchshöhe eines Pilzes angegeben in Zentimeter; hier könnte etwa $S_k = [0, 35]$ sein.

- Ordinale Merkmale: S_k besitzt eine im Kontext relevante (totale) Ordnung.

  Beispiel: Die Krümmung des Pilzhutes angegeben in der Form $S_k = \{$»kugelig«,»leicht gekrümmt«,»flach«$\}$.

- Nominale Merkmale: S_k besitzt keine relevante zusätzliche Struktur.

  Beispiel: Die Farbe des Pilzhutes;
  etwa $S_k = \{$»weiß«,»creme-farben«,»hellbraun«,»dunkelbraun«,»rot«$\}$.

Numerisch Merkmale werden auch als *quantitativ*, ordinale und nominale Merkmale als *qualitativ* bezeichnet.

Die Merkmale lassen sich in der Beschreibungsabbildung

$$\beta : \Omega \to S_1 \times \ldots \times S_m, \omega \mapsto (M_1(\omega), \ldots, M_m(\omega)) \tag{1}$$

zusammenfassen. Die Abbildung β ist im allgemeinen weder injektiv noch surjektiv: Einerseits können zwei verschiedene Objekte in den Ausprägungen aller Merkmalen M_k übereinstimmen und andererseits muss nicht jede Kombination von möglichen Ausprägungen durch ein Objekt realisiert werden.

Die Menge $S := S_1 \times \ldots \times S_m$ bezeichnet man als *Merkmalsraum* der Merkmale $M_1, \ldots, M_m$, wobei in die Definition eine bestimmte Nummerierung der Merkmale eingeht, die in der Praxis beliebig wählbar ist.

Eine *Datenmenge* ist eine Abbildung der Form

$$D : \{1, 2, \ldots, n\} \to S$$

mit einem beliebigen $n \in \mathbb{N}$; eine Datenmenge ist also entgegen der Erwartung im allgemeinen keine Menge. In der Literatur werden Datenmengen häufig als endliche Folge $(s^{(1)}, s^{(2)}, \ldots, s^{(n)})$ oft sogar unter Weglassen der Klammer als $s^{(1)}, s^{(2)}, \ldots, s^{(n)}$ angegeben. Obere Indizes werden hier deswegen verwendet, weil die Elemente $s^{(k)}$ selbst m-Tupel sind, deren Komponenten man mit unteren Indizes unterscheidet. Man beachte, dass man Datenmengen nicht einfach als Teilmengen von S definieren kann, da das mehrfache Auftreten eines Elements $s \in S$ in einer Datenmenge dann nicht erfasst wäre. Die Elemente $D(k)$ bzw. $s^{(k)}$ werden *Samples* genannt.

Im Data Mining werden aus der Menge Ω endliche Teilmengen $\Lambda \subseteq \Omega$ mehr oder weniger zufällig ausgewählt; man spricht auch von *Stichproben*[2]. Durch die Analyse der Eigenschaften der Objekte $\omega \in \Lambda$ soll Information über die Eigenschaften aller Objekte in Ω gewonnen werden.

Die Objekte in der Stichprobe werden üblicherweise nummeriert:

$$\Lambda = \{\omega_1, \omega_2, \ldots, \omega_n\}.$$

[2]Der Begriff »Stichprobe« wurde ursprünglich in der Eisenverhüttung, also dem Herstellungsprozess von Roheisen aus Eisenerzen mittels eines Hochofens, verwendet. In diesem Prozess werden zur Qualitätskontrolle regelmäßig Proben des flüssigen Eisens aus dem Hochofen entnommen. Hierzu wird dieser am sogenannten Stichloch geöffnet, ein Vorgang der als Abstich bezeichnet wird.

Die Einschränkung der Abbildung β auf Λ liefert dann die zu Λ gehörende Datenmenge

$$D : \{1, 2, \ldots, n\} \to S,\ k \mapsto \beta(\omega_k).$$

Datenmengen werden von Data Mining Software häufig in Form von Tabellen verarbeitet: Die Tabelle zu einer Datenmenge $D : \{1, 2, \ldots, n\} \to S$ besitzt dann üblicherweise $m + 1$ Spalten (m die Anzahl der Merkmale M_k) und die einzelnen Samples erscheinen als Zeilen der Tabelle, wobei in der ersten Spalte die Nummer des Samples und in den weiteren Spalten die Ausprägungen der Merkmale des Samples stehen.

Sind die Ausprägungen aller Merkmale Zahlen, so gibt man eine Datenmenge häufig auch in Form einer Matrix an:

$$\begin{pmatrix} s_1^{(1)} & s_2^{(1)} & \cdots & s_m^{(1)} \\ s_1^{(2)} & s_2^{(2)} & \cdots & s_m^{(2)} \\ \vdots & \vdots & \cdots & \vdots \\ s_1^{(n)} & s_2^{(n)} & \cdots & s_m^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times m},$$

wobei

$$\beta(\omega_k) = (s_1^{(k)}, s_2^{(k)}, \ldots, s_m^{(k)}).$$

Die in diesem Abschnitt beschriebene Situation wird im vorliegenden Skript als *Standard-Datenszenario* bezeichnet.

2 Das Klassifikationsproblem

Die Methoden des überwachten Data Mining dienen der Lösung zweier grundlegender Probleme, dem Klassifikations- und dem Regressionsproblem. Aus mathematischer Sicht können beide Probleme als Varianten des Problems der Funktionsapproximation aufgefasst werden. Wir beginnen mit der Diskussion des Klassifikationsproblems.

2.1 Naive Sicht

Motivation

Olivenöl setzt sich aus verschiedenen, in gebundener Form (»verestert«) vorliegenden Fettsäuren zusammen, die die Qualität des Öls bestimmen. Der Herstellungprozess von hochwertigem Olivenöl ist aufwendig und daher teuer. Die hohe Nachfrage nach solchen Ölen und die Preispolitik der Supermärkte hat zum Entstehen einer organisierten Kriminalität im Olivenölhandel geführt – siehe zum Beispiel [Gie]. Dabei werden hochwertige, italienische Olivenöle mit minderwertigen Ölen, Pflanzenölen anderer Art oder Olivensaft verschnitten/gepanscht. Diese Art von Betrug kann man zum Beispiel durch chemische Analyse der Zusammensetzung eines Olivenöls aufdecken, sofern diese Zusammensetzung charakteristisch für bestimmte Herkunftsregionen des Öls ist. Ein Vergleich der auf dem Flaschenetikett angegebenen Herkunft mit dem Ergebnis der chemischen Analyse einer Probe kann dann einen vorliegenden Betrug beweisen.

Abbildung 2: Olivenbäume und ihre Früchte

Die Frage, ob man die Herkunft eines Olivenöls anhand seiner chemischen Zusammensetzung ermitteln kann, wurde in der Arbeit [FAL] anhand von Daten zu 572 Olivenölproben unterschiedlicher Herkunft untersucht. Für jede Probe wurde der prozentuale Gehalt an acht verschiedenen Fettsäuren ermittelt. Die Proben selbst kamen aus neun Regionen Italiens, die zu drei Gebieten zusammengefasst wurden, nämlich Norditalien (Regionen Ostliguerien, Westligurien, Umbrien), Sardinien (Regionen Inland und Küste), sowie Süditalien (Regionen Nordapulien, Südapulien, Kalabrien, Sizilien).

Abbildung 3: Regionen Italiens
(Creative Commons Lizenz CC BY-SA 3.0)

Die Abbildung 4 zeigt die Zusammensetzung der Proben in Bezug auf Ölsäure und Linolsäure, sowie Linolensäure und Eikosensäure. Man kann erkennen, dass man durch kombinierte Betrachtung der Anteile dieser vier Fettsäuren jede Probe recht sicher dem jeweiligen Herkunftsgebiet zuordnen kann.

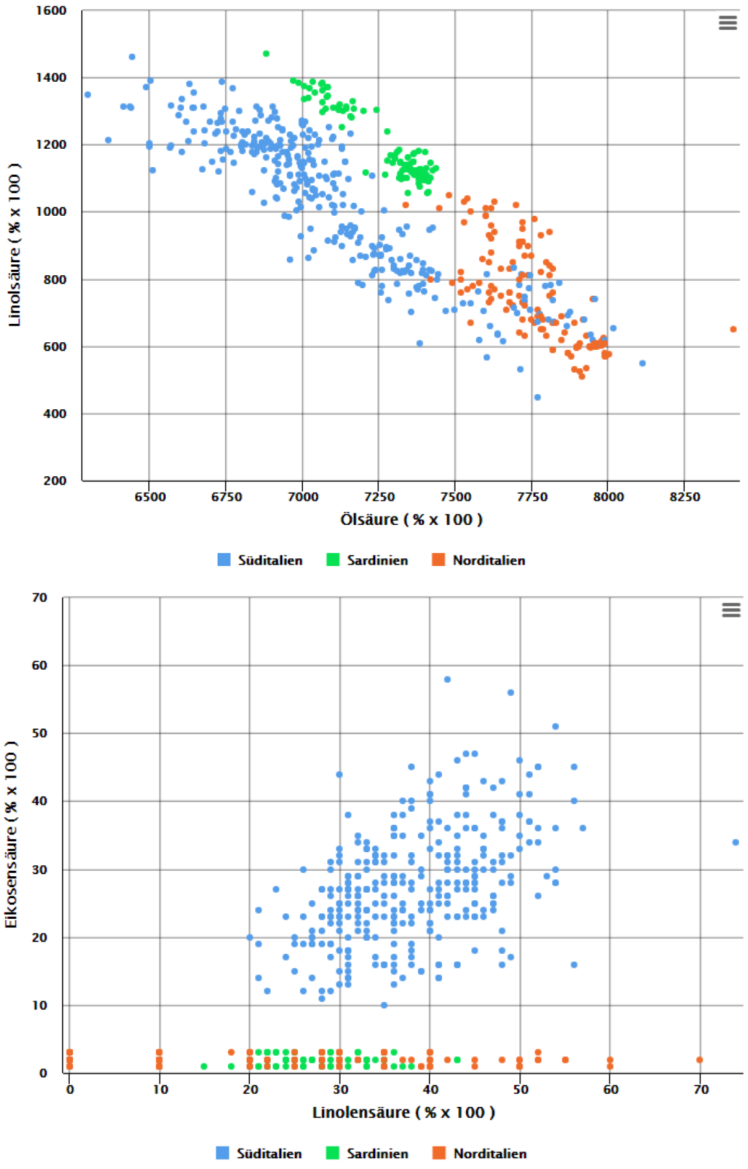


Abbildung 4: Chemische Zusammensetzung von Olivenölen

Für den praktischen Einsatz ist mathematisch formuliert also folgendes Problem zu lösen: Bestimme eine Abbildung

$$f : [0, 10000]^4 \to \{\text{Norditalien}, \text{Süditalien}, \text{Sardinien}\},$$

die nach Einsetzen der vier Werte $O(\omega), L_1(\omega), L_2(\omega), E(\omega)$ des Gehalts an Ölsäure, Linolsäure, Linolensäure und Eikosensäure einer Olivenölprobe ω mit hoher Treffsicherheit das Herkunftsgebiet $f(O(\omega), L_1(\omega), L_2(\omega), E(\omega))$ der Probe liefert. Wie die Notation bereits andeutet, werden also die Gehaltszahlen der vier genannten Fettsäuren als Ausprägungen von vier Merkmalen O, L_1, L_2, E von Olivenölproben aufgefasst. Der Definitionsbereich der Abbildung f kommt dadurch zustande, dass in den Daten der Fettsäuregehalt jeweils in »Prozentzahl mal 100«angegeben wird. Zur Bestimmung von f soll nur die Stichprobe Λ bestehend aus den 572 erwähnten Proben herangezogen werden. Die Abbildung f soll allerdings dennoch die Eigenschaft besitzen auch solchen chemischen Zusammensetzungen mit hoher Wahrscheinlichkeit die korrekte Herkunftsregion zuzuweisen, die nicht in der Stichprobe enthalten sind.

Ein Problem der gerade beschriebenen Art bezeichnet man als Klassifikationsproblem – es wird nun allgemein formuliert.

Das allgemeine Klassifikationsproblem

Prinzipiell geht man von dem im Abschnitt 1 beschriebenen Standard-Datenszenario aus. Allerdings wird zusätzlich angenommen, dass die Population Ω in $r \geq 2$ disjunkte Teilmengen zerfällt:

$$\Omega = \Omega_1 \cup \ldots \cup \Omega_r, \ \Omega_i \cap \Omega_j = \emptyset \text{ für } i \neq j.$$

Die Teilmengen Ω_i werden im Folgenden häufig als *Klassen* bezeichnet. Die Klassenzugehörigkeit kann als Merkmal der Objekte aufgefasst werden:

$$Y : \Omega \to \{1, 2, \ldots, r\}, \ \omega \mapsto k \text{ falls } \omega \in \Omega_k.$$

Das allgemeine Klassifikationsproblem kann jetzt vage wie folgt formuliert werden:

Klassifikationsproblem (vage): *Bestimme die Klassenzugehörigkeit $Y(\omega)$ eines beliebigen Objektes $\omega \in \Omega$ anhand der Ausprägungen der sonstigen gegebenen Merkmale dieses Objektes.*

Das Merkmal Y wird also als ein von den restlichen Merkmalen abhängiges behandelt, was sich in der Bezeichnungsweise niederschlagen sollte. Letztere werden daher im Unterschied zu der Darstellung im Abschnitt 1 mit $X_1, \ldots, X_p$ bezeichnet. Es gilt folglich $\{X_1, \ldots, X_p, Y\} = \{M_1, \ldots, M_m\}$ und der Merkmalsraum besitzt in der hier betrachteten allgemeinen Situation die Gestalt

$$S = S_1 \times \ldots \times S_p \times \{1, 2, \ldots, r\},$$

wobei S_k der Wertebereich von X_k ist.

Im Olivenölbeispiel ist $p = 4$ und die Merkmale $X_1, \ldots, X_4$ entsprechen den Merkmalen O, L_1, L_2, E. Weiter ist $r = 3$, wobei das Klassenmerkmal Y in diesem Fall anstelle der Ausprägungen $1, 2, 3$ die Ausprägungen »Norditalien«, »Süditalien«, »Sardinien« besitzt, was mathematisch keinen wesentlichen Unterschied macht.

Für das Weitere erweist es sich als praktisch die Menge

$$S_X := S_1 \times \ldots \times S_p$$

einzuführen, sowie die Abbildung:

$$X : \Omega \to S_X,\ \omega \mapsto (X_1(\omega), \ldots, X_p(\omega)).$$

Das Klassifikationsproblem lässt sich dann auf eine naive Weise präziser fassen:

Klassifikationsproblem (naiv): *Finde eine Abbildung* $f : S_X \to \{1, 2, \ldots, r\}$ *mit der Eigenschaft*

$$\forall \omega \in \Omega \quad f(X_1(\omega), \ldots, X_p(\omega)) = Y(\omega). \tag{2}$$

Dieses Zielsetzung ist jedoch aus zwei Gründen unrealistisch:

(UZ) Unvollständiger Zugriff: Der Datenanalytiker besitzt in aller Regel keinen vollständigen Zugriff auf die Population Ω.

(NDZ) Nicht deterministischer Zusammenhang: Die Existenz einer Abbildung f mit der Eigenschaft (2) impliziert, dass für zwei Objekte $\omega_1, \omega_2 \in \Omega$ mit der Eigenschaft

$$(X_1(\omega_1), \ldots, X_p(\omega_1)) = (X_1(\omega_2), \ldots, X_p(\omega_2))$$

stets auch

$$Y(\omega_1) = Y(\omega_2)$$

gilt, das heißt es muss ein deterministischer Zusammenhang zwischen den Ausprägungen der Merkmale $X_1, \ldots, X_p$ und denen des Merkmals Y bestehen. Ein Blick auf die Abbildung 4 zeigt, dass man dies nicht erwarten kann: Die chemischen Zusammensetzungen aus Ölsäure und Linolsäure von nord- und süditalienischen Olivenölen sind sich zum Teil so ähnlich, dass man trotz verschiedener Herkunft auch mit gleicher chemischer Zusammensetzung rechnen muss.

Dem Problem (UZ) wird im Data Mining durch die Betrachtung von Stichproben $\Lambda \subset \Omega$ begegnet. Die zugrundeliegende Idee dabei ist die folgende: Gilt die Gleichung (2) für »die meisten« $\omega \in \Lambda$ und ist die Stichprobe *repräsentativ*, das heißt spiegelt sie die wesentlichen Eigenschaften der Menge Ω in Bezug auf die Merkmale $X_1, \ldots, X_p, Y$ wider, so wird (2) auch für »die meisten« $\omega \in \Omega$ gelten. Insbesondere betrachtet man die durch die Klasseneinteilung von Ω induzierte Zerlegung

$$\Lambda = \Lambda_1 \cup \ldots \cup \Lambda_r, \ \Lambda_k := \Lambda \cap \Omega_k$$

der Stichproben Λ.

Um das Problem (NDZ) zu addressieren, wird die an die Abbildung f gestellte Forderung (2) dahingehend abgeschwächt, dass man sie nur noch für »möglichts viele« $\omega \in \Lambda$ fordert. Wir präzissieren dies: Die Menge aller Abbildungen $f : S_X \to \{1, 2, \ldots, r\}$ wird im Weiteren mit $\mathrm{Abb}(S_X, \{1, 2, \ldots, r\})$ bezeichnet. Solche Abbildungen nennt man im vorliegenden Kontext auch *Klassifikatoren*.

Definition 2.1: *Es liege die oben beschriebene, allgemeine Klassifikationsproblematik vor und $\Lambda \subset \Omega$ sei eine Stichprobe vom Umfang n. Die totale Trefferquote eines Klassifikators $f \in \mathrm{Abb}(S_X, \{1, 2, \ldots, r\})$ auf Λ ist dann definiert als:*

$$T(f, \Lambda) := \frac{t}{n}, \ t := |\{\omega \in \Lambda : f(X_1(\omega), \ldots, X_p(\omega)) = Y(\omega)\}|.$$

Entsprechend definiert man die Trefferquote von f in Klasse k auf Λ als

$$T_k(f, \Lambda) := \frac{t_k}{|\Lambda_k|}, \ t_k := |\{\omega \in \Lambda_k : f(X_1(\omega), \ldots, X_p(\omega)) = k\}|.$$

Eine weitere Präzissierung des Klassifikationsproblems ist dann:

Klassifikationsproblem (stichprobenabhängig): *Finde zu gegebener Stichprobe $\Lambda \subset \Omega$ einen Klassifikator $f : S_X \to \{1, 2, \ldots, r\}$ mit der Eigenschaft*

$$T(f, \Lambda) = \max(T(g, \Lambda) : g \in \mathrm{Abb}(S_X, \{1, 2, \ldots, r\})). \tag{3}$$

Bemerkungen:

1. Das Maximum in der Definition existiert, da als Trefferquoten nur die endlich vielen Werte $\frac{k}{n}$, $k \in \{0, \ldots, n\}$, in Frage kommen.

2. Die Funktion f ist nicht eindeutig bestimmt und hängt von der Stichprobe Λ ab. Daher ist es ein wesentliches Ziel bei der Datenanalyse diese Abhängigkeit zu quantifizieren.

3. Häufig möchte man nicht einfach die totale Trefferquote maximieren, sondern ist an einer gewichteten Berücksichtigung der klassenweisen Trefferquoten interessiert. Der Grund hierfür kann zum Beispiel sein, dass eine der Klassen in der Stichprobe besonders häufig vorkommt. Eine hohe Trefferquote in dieser Klasse zieht dann eine hohe totale Trefferquote nach sich. Das bedeutet aber, dass die Trefferquoten in den kleineren Klassen vernachlässigt werden. Durch ein Hochgewichten der Trefferquoten der kleineren Klassen kann man diesen Effekt ausgleichen.

   Formal führt man Gewichte $w_1, \ldots, w_r \in [0, 1]$ mit $w_1 + w_2 + \ldots + w_r = 1$ ein und betrachtet dann

   $$T_w(g, \Lambda) := w_1 T_1(g, \Lambda) + w_2 T_2(g, \Lambda) + \ldots + w_r T_r(g, \Lambda) \tag{4}$$

   anstelle der totalen Trefferquote $T(g, \Lambda)$.

Nächste-Nachbarn-Klassifikation

Es stellt sich nun die Frage, wie man das Optimierungsproblem (3) löst. Beim Betrachten der Abbildung 4 beobachtet man Folgendes: Objekte ω und ω' gehören häufig dann zur gleichen Klasse Ω_k, wenn ihre Merkmalsausprägungen $X_1(\omega), \ldots, X_p(\omega)$ und $X_1(\omega'), \ldots, X_p(\omega')$ ähnlich zueinander sind. Um diesen Gedanken weiter zu verfolgen, muss man also Elemente $x \in S_X$ miteinander vergleichen können, und zwar in einer Art und Weise, die dem vorliegenden Anwendungskontext entspricht. Für das Weitere wird daher die folgende Annahme gemacht:

Auf S_X liege eine Metrik $d : S_X \times S_X \to \mathbb{R}$ vor, die zum Anwendungskontext passt, in diesem also insbesondere interpretierbar ist.

Naheliegend ist dann die folgende algorithmische Definition einer Lösung f von (3):

1. Man wähle ein Zahl $R > 0$; die Bedeutung dieser Zahl wird im folgenden Schritt erkennbar.

2. Zur Bestimmung von $f(x)$ betrachte man die abgeschlossene Kugel

   $$B[x, R] := \{x' \in S_X : d(x, x') \leq R\}$$

   mit Mittelpunkt x und Radius R in dem metrischen Raum (S_X, d) und bestimme die Zahlen

   $$n_\ell := |\{\omega \in \Lambda_\ell : X(\omega) \in B[x, R]\}|.$$

3. Man setze $f(x) := L$, falls

   $$n_L = \max(n_1, n_2, \ldots, n_r)$$

   gilt. Ist L durch diese Eigenschaft nicht eindeutig bestimmt, so wähle man unter den möglichen Indizes einen zufällig aus.

Der beschriebene Algorithmus besitzt die Schwäche, dass die Kugel $B[x, R]$ möglicherweise für manche x keine Elemente von $X(\Lambda)$, also der zu Λ gehörenden Datenmenge, enthält. In diesem Fall kann der Funktionswert $f(x)$ nicht bestimmt werden. Dieses Problem könnte man durch Wahl eines hinreichend großen Radius' beheben. Andererseits sollte allerdings R eher »klein« sein, da man nur nächste Nachbarn von x zur Bestimmung der Klasse heranziehen möchte. Daher und aus statistischen Gründen modifiziert man den Algorithmus auf die im folgenden beschriebene Weise: Statt einen Kugelradius R vorzugeben, gibt man die Anzahl der Elemente der Datenmenge vor, die in die Berechnung des Funktionswerts $f(x)$ eingehen sollen. Diese sollten natürlich alle möglichst nahe bei x liegen, was zu folgender Definition führt:

DEFINITION 2.2: *Das oben beschriebene allgemeine Klassifikationsproblem liege vor und es sei $x_1, x_2, \ldots, x_n$ die Datenmenge zu einer Stichprobe $\Lambda \subseteq \Omega$.*

Die Elemente $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ heißen nächste Nachbarn des Elements $x \in S_X$, falls für jedes $i \notin \{i_1, \ldots, i_k\}$ die Bedingung

$$\forall j \in \{1, 2, \ldots, k\} \quad d(x_{i_j}, x) \leq d(x_i, x)$$

erfüllt ist.

Der modifizierte Algorithmus ist in der folgenden Definition enthalten:

DEFINITION 2.3: *Das oben beschriebene allgemeine Klassifikationsproblem liege vor und $\Lambda \subseteq \Omega$ sei eine Stichprobe. Zu jeder Zahl $k \in \mathbb{N}$ ist ein k-Nächste-Nachbarn-Klassifikator (k-NN-Klassifikator)*

$$f : S_X \to \{1, 2, \ldots, r\}, \; x \mapsto f(x)$$

wie folgt definiert:

1. *Zu $x \in S_X$ wähle man k nächste Nachbarn $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ in der Datenmenge zur Stichprobe Λ.*

2. *Man bestimme die Zahlen*

$$n_\ell := |\{j \in \{1, \ldots, k\} : x_{i_j} \in X(\Omega_\ell)\}|.$$

3. *Man setze $f(x) := L$, falls*

$$n_\ell = \max(n_1, n_2, \ldots, n_r)$$

gilt. Ist L durch diese Eigenschaft nicht eindeutig bestimmt, so wähle man unter den möglichen Indizes einen zufällig aus.

Es stellt sich unmittelbar die Frage nach der »richtigen« Wahl von k. Die Antwort ist in vielen Fällen sehr einfach:

FESTSTELLUNG 2.4: *Besitzt die Stichprobe Λ die Eigenschaft*

$$\forall \omega, \omega' \in \Lambda \quad X(\omega) = X(\omega') \Rightarrow Y(\omega) = Y(\omega'),$$

so gilt für jeden 1-NN-Klassifikator f: $T(f, \Lambda) = 1$; insbesondere löst f das Optimierungsproblem (3).

BEWEIS: Es sei $x = X(\omega)$ mit $\omega \in \Lambda$. Nach Voraussetzung gibt es kein $\omega' \in \Lambda$ mit $Y(\omega) \neq Y(\omega')$. Folglich besitzt jeder 1-Nächste Nachbar von x diesselbe Klassenzugehörigkeit L wie ω, womit $f(x) = L$ ist. □

Die obige Feststellung könnte man als Indiz dafür betrachten, dass der Zielansatz (3) der richtige ist. Die folgende Diskussion wird aber zeigen, dass dies nicht der Fall ist.

DISKUSSION EINES EXPERIMENTS

Jeder basierend auf einer Stichprobe $\Lambda \subseteq \Omega$ ermittelte Klassifikator $f : S_X \to \{1, 2, \ldots, r\}$ soll auch Objekte $\omega \in \Omega \setminus \Lambda$ möglichst oft korrekt klassifizieren:

$$f(X(\omega)) = Y(\omega).$$

Diese Eigenschaft von f bezeichnet man als *Generalisierungsfähigkeit*. In Bezug auf diese Eigenschaft wurde folgendes Experiment durchgeführt: In dem Experiment wurde die früher beschriebene Stichprobe Λ aus 572 chemischen Zusammensetzungen von Olivenölen aus 3 Herkunftsregionen zufällig in zwei disjunkte Teilmengen geteilt:

$$\Lambda = \Lambda_{\text{train}} \cup \Lambda_{\text{test}}. \tag{5}$$

Die im Weiteren als *Trainingsmenge* bezeichnete Teilmenge Λ_{train} enthielt 60% der Daten, die *Testmenge* Λ_{test} entsprechend 40%. Basierend auf der Menge Λ_{train} wurden k-NN-Klassifikatoren f_k für $k \in \{1, 2, \ldots, 13\}$ erstellt, wobei nur die Merkmale »Ölsäuregehalt« und »Linolsäuregehalt« benutzt wurden. Die totalen Trefferquoten $T(f_k, \Lambda_{\text{train}})$ auf der Trainingsmenge und $T(f_k, \Lambda_{\text{test}})$ auf der Testmenge wurden für alle Klassifikatoren f_k ermittelt. Die Trefferquoten $T(f_k, \Lambda_{\text{test}})$ können dabei als Maß für die Generalisierungsfähigkeit von f_k gesehen werden.

Um einen Eindruck von den im Experiment ermittelten Klassifikatoren zu erhalten, ist der Klassifikator f_5 in den Abbildungen 5 und 6 grafisch dargestellt. Man beachte hierbei, dass die Ausprägungen der betrachteten Merkmale *vor* der Bestimmung des Klassifikators standardisiert wurden.

Das Experiment wurde vier mal mit unterschiedlichen zufälligen Aufteilungen (5) wiederholt. Die Ergebnisse für die verschiedenen Aufteilungen sind in Abbildung 7 in unterschiedlichen Farben dargestellt. Die durchgezogenen Linien stellen dabei die Verläufe der Trefferquoten $T(f_k, \Lambda_{\text{train}})$ abhängig von k und die gestrichelten Linien die der Trefferquoten $T(f_k, \Lambda_{\text{test}})$ dar.

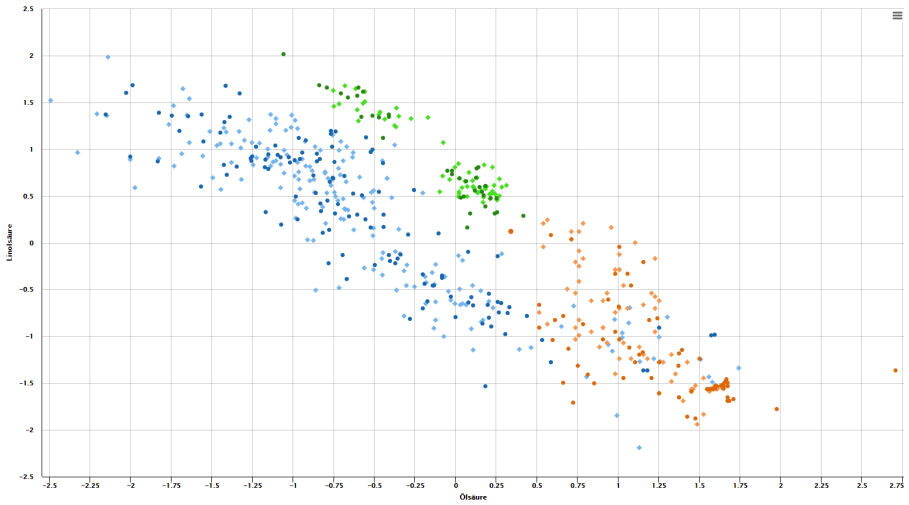


Abbildung 5: 5-NN-Klassifikator für die Datenmenge Olive Oils
Hell gefärbte Punkte: Trainingsmenge mit wahrer Klassenzugehörigkeit
Dunkel gefärbte Punkte: Testmenge mit prognostizierter Klassenzugehörigkeit
Farbkodierung der Regionen wie in Abbildung 4

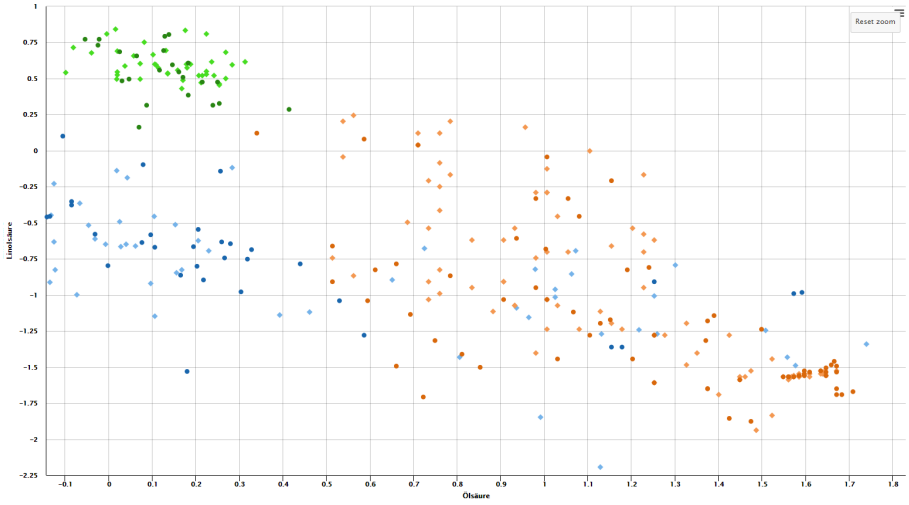


Abbildung 6: 5-NN-Klassifikator für die Datenmenge Olive Oils (Ausschnitt)
Farbkodierung der Punkte wie in Abbildung 5
Farbkodierung der Regionen wie in Abbildung 4

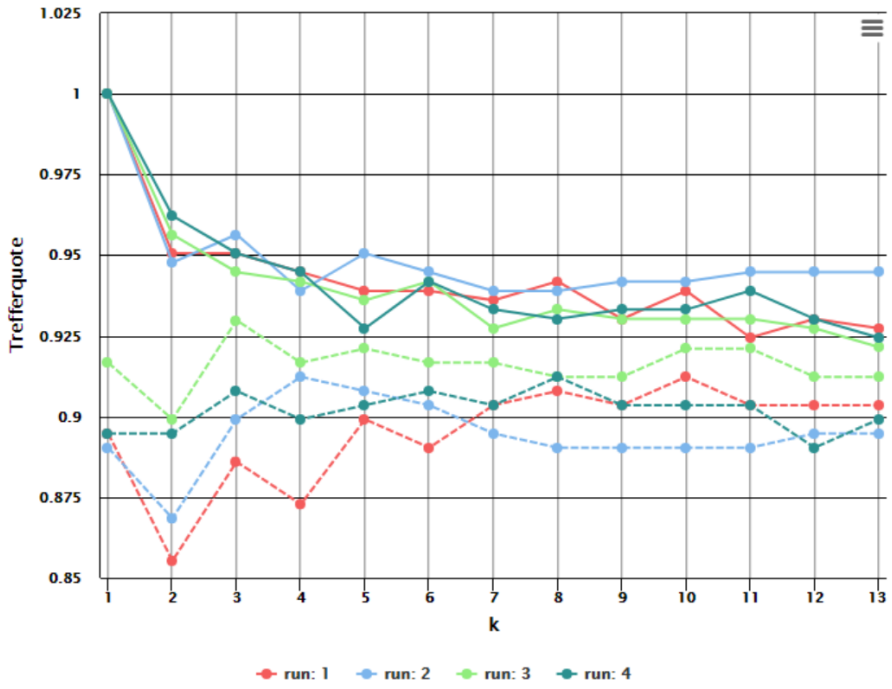


Abbildung 7: Trefferquoten von k-NN-Klassifikatoren abhängig von k: durchgezogene Linien: $T(f_k, \Lambda_{\text{train}})$, gestrichelte Linien: $T(f_k, \Lambda_{\text{test}})$

Folgende Beobachtungen kann man anhand der Ergebnisse des Experimentes machen:

- Es gilt stets $T(f_k, \Lambda_{\text{train}}) > T(f_k, \Lambda_{\text{test}})$.

  Interpretation: Die Trefferquote auf der Trainingsmenge ist deswegen höher als auf der Testmenge, weil die in der Trainingsmenge steckende Information über die Klassenaufteilung der Objekte ω in die Definition von f_k eingeht.

- Wie erwartet – siehe Feststellung 2.4, ist $T(f_1, \Lambda_{\text{train}}) = 1$. Bemerkenswert ist aber, dass der Mittelwert der vier Werte von $T(f_1, \Lambda_{\text{test}})$ bei etwa 0.898 liegt, während der Mittelwert der vier Werte für $T(f_5, \Lambda_{\text{test}})$ bei etwa 0.907 liegt, also größer ist. Dies gilt auch für andere Werte von k, wie etwa $k = 7$.

Interpretation: Wenn man die Mittelwerte der Trefferquoten auf der Testmenge als Maß für die Generalisierungsfähigkeit nutzt, so besitzt der Klassifikator mit der höchsten Trefferquote auf der Trainingsmenge nicht notwendig die höchste Generalisierungsfähigkeit. Dies kann an einem Sachverhalt liegen, den man als *Overfitting* bezeichnet: Der Klassifikator f_1 beispielsweise ist aufgrund seiner Definition präzise an die Trainingsstichprobe angepasst. Die Samples dieser Stichprobe können jedoch auch nicht typische oder Grenzfälle enthalten, also etwa Olivenölproben, die zwar definitiv sardischer Herkunft von ihrer Zusammensetzung aber eher vom norditalienischen Typ sind. Die präzise Anpassung an solche Fälle führt in der Teststichprobe zu einer erhöhten Fehlerrate. Man beachte, dass im vorliegenden Fall Klassifikatoren f_k mit $k > 1$ weniger zu einer solchen präzisen Anpassung neigen.

Wenn man nur Einzelverläufe betrachtet, wie zum Beispiel die in hellgrüner Farbe dargestellte dritte Wiederholung des Experiments, so ist der Effekt noch ausgeprägter zu sehen: Es ist dann $T(f_1, \Lambda_{\text{test}}) = 0.916$ aber $T(f_3, \Lambda_{\text{test}}) = 0.929$ und $T(f_5, \Lambda_{\text{test}}) = T(f_{10}, \Lambda_{\text{test}}) = T(f_{11}, \Lambda_{\text{test}}) = 0.920$.

Schlussfolgerung: Da die Generalisierungsfähigkeit eines Klassifikators eine wichtige Eigenschaft ist, ist (3) immer noch keine angemessene Formulierung des Ziels, das bei der Lösung des allgemeinen Klassifikationsproblems verfolgt werden sollte.

2.2 Stochastische Sicht

Der im letzten Abschnitt eingeführte k-Nächste-Nachbarn-Klassifikator wurde basierend auf empirischen Überlegungen eingeführt. Das ist im Data Mining ein durchaus akzeptiertes Prinzip. Weiter- oder tiefergehende Fragen kann man auf diese Weise jedoch nicht beantworten. Solche sind zum Beispiel:

- Welche Eigenschaften zeichnen k-Nächste-Nachbarn-Klassifikatoren im Vergleich mit anderen Klassifikatoren aus?
- Welche datenanalytischen Stärken und Schwächen besitzen k-Nächste-Nachbarn-Klassifikatoren?

Eine präzisere mathematische Formulierung des Klassifikationsproblems ist erforderlich. Wegen der Abhängigkeit der Ergebnisse von zufällig aus Ω gezogenen Stichproben, muss hierzu die Stochastik genutzt werden. Im Folgenden werden die notwendigen stochastischen Fakten zusammengetragen.

1. **Ereignisse bei der Datenerhebung:** Im Zusammenhang mit der Bestimmung der Ausprägungen von Merkmalen von Objekten $\omega \in \Omega$ treten bestimmte Ereignisse auf, die im Kontext des Data Mining relevant sind. Ist beispielsweise Ω die Gesamtheit aller im Jahr 2020 auf den Markt gebrachten Flaschen von italienischem Olivenöl, so sind solche Ereignisse:

   »Aus Ω wird bei einer Stichprobe eine Flasche mit einem Ölsäuregehalt $O(\omega) \in [72\%, 73\%]$ gezogen.«

   »Eine Stichprobe $\Lambda \subseteq \Omega$ enthält 5 Flaschen Olivenöl aus dem sardischen Inland.«

   Alle diese relevanten Ereignisse werden in einer σ-Algebra $\mathcal{E}$ über Ω zusammengefasst.

   Man beachte, dass diese σ-Algebra Bedingungen allgemeiner Art, sowie kontextspezifische Bedingungen erfüllen muss: Allgemein muss zum Beispiel $\Omega_k \in \mathcal{E}$ gelten, da das Ereignis »Bei einer Stichprobe wird ein Objekt aus der Klasse k gezogen.« immer relevant ist. Andererseits kann man den Fettsäuregehalt nicht genau messen, sondern abhängig von der eingesetzten Messtechnik nur bis auf einen unvermeidbaren Messfehler. Folglich ist das Ereignis »Bei einer Stichprobe wird eine Flasche Olivenöl mit einem Ölsäuregehalt $O(\omega) = 72,976318\%$ gezogen.« nicht relevant, falls der Messfehler größer als 0.01 sein kann.

2. **Ereigniswahrscheinlichkeiten:** Für das Eintreten der Ereignisse $E \in \mathcal{E}$ existieren Wahrscheinlichkeiten $P(E)$. Diese liefern ein Wahrscheinlichkeitsmaß $P : \mathcal{E} \to [0, 1]$.

   Das Wahrscheinlichkeitsmaß P ist unbekannt.

3. **Priors:** Insbesondere existieren die unbekannten Wahrscheinlichkeiten

   $$p_k := P(\Omega_k) = P(Y(\omega) = k);$$

   diese nennt man *a-priori Wahrscheinlichkeiten* oder *Priors*.

4. **Ereigniswahrscheinlichkeiten bei gegebener Klasse:** Das Wahrscheinlichkeitsmaß $P : \mathcal{E} \to [0,1]$ liefert klassenweise Wahrscheinlichkeitsmaße $P_k : \mathcal{E}_k \to [0,1]$: Für jedes $k \in \{1,2,\ldots,r\}$ betrachtet man die σ-Algebra
$$\mathcal{E}_k := \{E \cap \Omega_k : E \in \mathcal{E}\}$$
und definiert das Wahrscheinlichkeitsmaß P_k durch die Formel
$$P_k(E') := \frac{P(E')}{P(\Omega_k)} = \frac{1}{p_k} P(E').$$

5. **Mischverteilung:** Für das Maß P gilt nach dem Satz von der totalen Wahrscheinlichkeit
$$P(E) = \sum_{k=1}^{r} p_k P_k(E \cap \Omega_k).$$

6. **Bedingte Wahrscheinlichkeiten:** Die Wahrscheinlichkeiten $P_k(E')$ kann man auch als bedingte Wahrscheinlichkeiten deuten: Allgemein ist
$$P(E|\Omega_k) := \frac{P(E \cap \Omega_k)}{P(\Omega_k)},$$
die Wahrscheinlichkeit für das Eintreten des Ereignisses E gegeben, dass das Ereignis Ω_k eingetreten ist.

Alle bisher eingeführten mathematischen Objekte sind unbekannt, müssen also basierend auf der Datenmenge zu einer Stichprobe $\Lambda \subseteq \Omega$ geschätzt werden. Da nur die Ausprägungen der Merkmale $X_1, \ldots, X_p, Y$ beobachtbar sind, müssen diese entsprechend in die stochastische Sicht einbezogen werden. Dies geschieht, indem man sie als Zufallsvariablen betrachtet.

7. **Zufallsvariablen:** Ein Paar $(M, \mathcal{M})$ bestehend aus einer Menge M und einer σ-Algebra $\mathcal{M}$ über M heißt *Messraum*.

   Eine *M-wertige Zufallsvariable* ist eine messbare Abbildung $Z : \Omega \to M$, das heißt Z besitzt die Eigenschaft
$$\forall N \in \mathcal{M} \quad Z^{-1}(N) \in \mathcal{E}.$$
   Die Urbildmenge $Z^{-1}(N) = \{\omega \in \Omega : Z(\omega) \in N\}$ schreibt man häufig in der intuitiven Kurzform $\{Z \in N\}$.

Von besonderer Bedeutung sind *reelle Zufallsvariablen*: Über den reellen Zahlen $\mathbb{R}$ ist der *Borelkörper* $\mathcal{B}^1$ eine natürliche σ-Algebra. $\mathcal{B}^1$ ist die kleinste σ-Algebra, die alle offenen Intervalle in $\mathbb{R}$ enthält. Eine reelle Zufallsvariable ist eine Borel-messbare Funktion $Z : \Omega \to \mathbb{R}$.

8. **Merkmale als Zufallsvariablen:** Die Merkmale $X_1, \ldots, X_p, Y$ sind bei einer dem jeweils vorliegenden Anwendungskontext angepassten Wahl der σ-Algebra $\mathcal{E}$ und von σ-Algebren $\mathcal{S}_k$ auf S_k fast automatisch Zufallsvariablen. Die folgenden Beispiele zeigen, was das bedeutet:

   Ist $X_k : \Omega \to S_k$ ein kategorielles oder ordinales Merkmal mit endlich vielen Ausprägungen $s \in S_k$, so wird man das Eintreten von »Aus Ω wird ein Objekt ω mit einer der Ausprägungen $s_1, \ldots, s_q \in S_k$ von X_k gezogen.« als relevantes Ereignis ansehen. Folglich sollte

   $$\{\omega \in \Omega : X_k(\omega) \in \{s_1, \ldots, s_q\}\} = X_k^{-1}(\{s_1, \ldots, s_q\}) \in \mathcal{E}$$

   gelten. Weiter wird X_k messbar, wenn man für $\mathcal{S}_k$ die Potenzmenge $\mathcal{P}(S_k)$ als σ-Algebra wählt.

   Sei nun $X_k : \Omega \to S_k \subseteq \mathbb{R}$ ein reellwertiges Merkmal, dessen Ausprägungen durch Messung mit einem Messgerät ermittelt werden. Da solche Geräte nicht völlig präzise arbeiten, ist eine Angabe der Form $X_k(\omega) = s$ nicht sinnvoll. Besser ist es Angaben der Form $X_k(\omega) = s \pm \epsilon$ mit einem $\epsilon > 0$ zu betrachten, also in die σ-Algebra $\mathcal{E}$ Ereignisse der Form »Aus Ω wird ein Objekt ω mit der Eigenschaft $X_k(\omega) \in (s - \epsilon, s + \epsilon)$ gezogen.« aufzunehmen. Ist nun S_k selbst ein Intervall, so wird X_k messbar, wenn man über S_k die σ-Algebra

   $$\mathcal{S}_k := \{B \cap S_k : B \in \mathcal{B}^1\}$$

   wählt.

9. **Wahrscheinlichkeiten auf dem Merkmalsraum:** Die Merkmale $X_1, \ldots, X_p$ können zu einer Zufallsvariablen

   $$X : \Omega \to S_X,\ \omega \mapsto (X_1(\omega), \ldots, X_p(\omega))$$

   zusammengefasst werden, indem man auf dem kartesischen Produkt $S_X = S_1 \times \ldots \times S_p$ die Produkt-σ-Algebra

   $$\mathcal{S} := \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_p$$

wählt. Letztere ist die von den Mengen $A_1 \times \cdots \times A_p$, $A_k \in \mathcal{S}_k$, erzeugte σ-Algebra.

Man kann nun das Bildmaß

$$P_X : S_X \to [0,1],\ A \mapsto P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

als Wahrscheinlichkeitsmaß auf S_X einführen.

10. **A-posteriori-Wahrscheinlichkeiten:** Das Klassifikationsproblem besteht grob darin, die Klassenzugehörigkeit $Y(\omega)$ eines Objekts ω anhand der Merkmalsausprägungen $(X_1(\omega), \ldots, X_p(\omega)) = X(\omega)$ zu bestimmen. Es ist klar, dass dabei die bedingten Wahrscheinlichkeiten

$$P(Y = k | X \in A),\ A \in \mathcal{S},$$

also die Wahrscheinlichkeit dafür, dass ein Objekt zur Klasse k gehört gegeben, dass seine sonstigen Merkmalsausprägungen in einem bestimmten Bereich A liegen, eine wichtige Rolle spielen. Für diese gilt

$$P(Y = k | X \in A) = P(\Omega_k)\frac{P(X \in A | Y = k)}{P(X \in A)} = p_k \frac{P(X \in A | Y = k)}{P_X(A)}$$

für jedes $k \in \{1, \ldots, r\}$.

Es fehlt nun nur noch, die Klassifikatoren selbst in die stochastische Sicht zu integrieren: In der naiven Sicht des Problems wurden als Klassifikatoren *alle* Abbildungen $f : S_X \to \{1, 2, \ldots, r\}$ zugelassen. Das ist bei der stochastischen Herangehensweise nicht möglich, wie gleich klar wird.

11. **Suchraum für Klassifikatoren:** Um die Güte eines Klassifikators $f : S_X \to \{1, 2, \ldots, r\}$ zu quantifizieren werden im Folgenden die Wahrscheinlichkeiten von Trefferereignissen

$$\{\omega \in \Omega : f(X_1(\omega), \ldots, X_p(\omega)) = Y(\omega)\}$$

betrachtet. Damit dies möglich ist, muss die Abbildung

$$f \circ X : \Omega \to \{1, 2 \ldots, r\},\ \omega \mapsto f(X_1(\omega), \ldots, X_p(\omega))$$

als Zufallsvariable betrachtet werden, was wiederum im allgemeinen nur möglich ist, wenn $f : S_X \to \{1, 2, \ldots, r\}$ messbar ist. Dies ist genau dann der Fall, wenn $f^{-1}(k) \in \mathcal{S}$ für jedes $k \in \{1, 2, \ldots, r\}$ gilt.

Als Konsequenz muss man »gute« Klassifikatoren in der Teilmenge

$$\mathrm{M}(S_X, \{1, 2, \ldots, r\}) \subseteq \mathrm{Abb}\,(S_X, \{1, 2, \ldots, r\})$$

der messbaren Abbildungen suchen. Wie sich im Lauf der Vorlesung zeigen wird, kann es allerdings durchaus sinnvoll sein den Suchraum für einen »guten« Klassifikator von vornherein auf eine Teilmenge

$$\mathbf{F} \subseteq \mathrm{M}(S_X, \{1, 2, \ldots, r\})$$

zu beschränken. Die Menge $\mathbf{F}$ wird auch als *Modellraum* bezeichnet.

Mit Hilfe der bis zu diesem Punkt eingeführten stochastischen Begriffe kann man eine Optimalitätseigenschaft von Klassifikatoren präzise formulieren:

Definition 2.5: *Es liege das in den Punkten 1 bis 11 formulierte Klassifikationsszenario vor. Für einen messbaren Klassifikators $f : S_X \to \{1, \ldots, r\}$ wird dann die Größe*

$$P(f(X_1, \ldots, X_p) = Y) = P(\{\omega \in \Omega : f(X_1(\omega), \ldots, X_p(\omega)) = Y(\omega)\})$$

als totale Trefferwahrscheinlichkeit bezeichnet.

Entsprechend bezeichnet man die Größe

$$P(f(X_1, \ldots, X_p) = k | Y = k) = \frac{1}{p_k} P(f(X_1, \ldots, X_p) = k \wedge Y = k)$$

als Trefferwahrscheinlichkeit in der Klasse k.

Sind $w_1, \ldots, w_r \in [0, 1]$ Gewichte, so bezeichnet man die Größe

$$P(f(X_1, \ldots, X_p) = Y)_w := \sum_{k=1}^{r} w_k P(f(X_1, \ldots, X_p) = k | Y = k)$$

als gewichtete totale Trefferwahrscheinlichkeit.

Bemerkungen:

1. Da f und damit $f \circ X$ messbar ist, liegt für jedes k die Menge

   $$(f \circ X)^{-1}(k) \cap \Omega_k$$

   in der σ-Algebra $\mathcal{E}$. Damit liegt auch

   $$\{\omega \in \Omega : f(X_1(\omega), \ldots, X_p(\omega)) = Y(\omega)\} = \bigcup_{k=1}^{r} (f \circ X)^{-1}(k) \cap \Omega_k$$

   in $\mathcal{E}$, die totale Trefferwahrscheinlichkeit ist also wohldefiniert. Die Mengen rechterhand sind paarweise disjunkt, daher folgt:

2. $P(f(X_1, \ldots, X_p) = Y) = \sum_{k=1}^{r} P(X \in f^{-1}(k) \wedge Y = k)$.

3. Nach Punkt 5 des Klassifikationsszenarios gilt

$$P(f(X_1, \ldots, X_p) = Y)_w = P(f(X_1, \ldots, X_p) = Y),$$

falls man $w_k = p_k$ wählt.

Mit Hilfe der Trefferwahrscheinlichkeit(en) kann man nun auch präzise definieren, was man unter einer Lösung des Klassifikationsproblems verstehen möchte:

DEFINITION 2.6: *Es liege das in den Punkten 1 bis 11 formulierte Klassifikationsszenario vor. Weiter seien* $\mathbf{F} \subseteq \mathrm{M}(S_X, \{1, 2, \ldots, r\})$ *eine Menge messbarer Klassifikatoren und* $w_1, \ldots, w_r \in [0, 1]$ *Gewichte. Als allgemeines Klassifikationsproblem bezeichnet man das Optimierungsproblem*

$$\mathrm{argmax}(P(f(X_1, \ldots, X_p) = Y)_w : f \in \mathbf{F}). \tag{6}$$

Jede Lösung dieses Optimierungsproblems heißt entsprechend Lösung des Klassifikationsproblems.

BEZEICHNUNGEN: In dem durch die Definition 2.6 gegebenen Rahmen sind in der Literatur verschiedene Bezeichnungen für die Zufallsvariablen $X_1, \ldots, X_n, Y$ gebräuchlich: Die X_i werden als unabhängige Variablen, Inputs, Prädiktoren oder einfach als Merkmale bezeichnet. Die Zufallsvariable Y wird entsprechend abhängige Variable, Output oder Response genannt.

3 Bayes-Klassifikation

Basierend auf den theoretischen Überlegungen des vorigen Abschnitts werden in diesem Abschnitt explizite Lösungen des Klassifikationsproblems in zwei Fällen gewonnen, in denen jeweils weitere Annahmen über die Merkmale $X_1, \ldots, X_p$ gemacht werden. Die erhaltenen Ergebnisse gelten auch ohne diese Zusatzannahmen, was hier aber nicht bewiesen wird.

3.1 Inputs mit abzählbar vielen Ausprägungen

Im Folgenden wird angenommen, dass jedes Merkmal X_k abzählbar viele Ausprägungen besitzt; man kann also annehmen, dass die Menge S_k jweils endlich oder abzählbar unendlich ist. In diesem Fall ist auch die Menge S_X abzählbar, da endliche Produkte abzählbarer Mengen selbst abzählbar sind. Wie im vorherigen Abschnitt ausgeführt, wird S_k dann mit der Potenzmenge $\mathcal{P}(S_k)$ als σ-Algebra versehen. Für die Produkt-σ-Algebra auf S_X folgt

$$\mathcal{S} = \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_p = \mathcal{P}(S_X).$$

Die gewichtete totale Trefferwahrscheinlichkeit einer messbaren Funktion $f : S_X \to \{1, 2, \ldots, r\}$ lässt sich dann als

$$\begin{aligned}
P(f(X_1, \ldots, X_p) = Y)_w &= \sum_{k=1}^{r} w_k P(f(X_1, \ldots, X_p) = k | Y = k) \\
&= \sum_{k=1}^{r} w_k P(X \in f^{-1}(k) | Y = k) \\
&= \sum_{k=1}^{r} \tfrac{w_k}{p_k} P(X \in f^{-1}(k) \wedge Y = k) \\
&= \sum_{k=1}^{r} \tfrac{w_k}{p_k} \sum_{x \in f^{-1}(k)} P(X = x \wedge Y = k) \\
&= \sum_{k=1}^{r} \tfrac{w_k}{p_k} \sum_{x \in f^{-1}(k)} P(Y = k | X = x) P_X(x) \\
&= \sum_{k=1}^{r} \sum_{x \in f^{-1}(k)} \tfrac{w_k}{p_k} P(Y = k | X = x) P_X(x).
\end{aligned}$$

schreiben, wobei die inneren Summen gegebenenfalls absolut konvergente, unendliche Reihen sind. Die Summe beziehungsweise Reihe auf der rechten Seite der letzten Gleichung besitzt die Eigenschaften:

- sämtliche Summanden sind nicht-negativ,

- jedes $x \in S_X$ legt genau einen Summanden fest,
- $P_X(x)$ ist nicht von f abhängig.

Die Summe wird folglich maximal, falls f so gewählt wird, dass jeder der Terme $\frac{w_k}{p_k} P(Y = k|X = x)$ maximal wird. Man beachte dabei, dass ein Klassifikator f durch die Partition

$$S_X = f^{-1}(1) \cup f^{-1}(2) \cup \ldots \cup f^{-1}(r)$$

eindeutig bestimmt ist. Die gerade durchgeführte Überlegung beweist den nachfolgenden Satz:

SATZ 3.1: *Es liege das in den Punkten 1 bis 11 von Abschnitt 2.2 formulierte Klassifikationsszenario vor, wobei die Mengen $S_1, \ldots, S_p$ zu den Merkmalen $X_1, \ldots, X_p$ jeweils abzählbar seien.*

Weiter sei $\mathbf{F} = \mathrm{M}(S_X, \{1, 2, \ldots, r\})$ und $w_1, \ldots, w_r \in [0, 1]$ seien beliebige Gewichte.

Dann ist die Abbildung

$$\begin{array}{rcl} b : S_X & \to & \{1, \ldots, r\} \\ b(x) & := & k, \\ \text{wobei } \frac{w_k}{p_k} P(Y = k|X = x) & = & \max(\frac{w_j}{p_j} P(Y = j|X = x) : j \in \{1, \ldots, r\}) \end{array} \tag{7}$$

eine Lösung des Optimierungsproblems (6).

Man nennt b einen Bayes[3]-Klassifikator zu dem gegebenen Klassifikationsproblem.

BEMERKUNG: Das Wahrscheinlichkeitsmaß P und die Gewichte $w_1, \ldots, w_r$ legen b nicht eindeutig fest, da das Maximum in (7) für mehrere Indizes k angenommen werden kann. In diesem Fall kann man theoretisch einen beliebigen dieser Indizes k als Funktionswert $b(x)$ wählen. In der Praxis ist das je nach Kontext nicht unbedingt empfehlenswert – siehe das Beispiel 3.2.

In der Praxis kann der Bayes-Klassifikator b (natürlich) nicht exakt angegeben werden, da P unbekannt ist. Stattdessen ermittelt man eine Schätzung $\hat{b}$ von b anhand einer Stichprobe $\Lambda \subseteq \Omega$: Ist die zu Λ gehörende Datenmenge

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}),\ x^{(i)} \in S_X, y^{(i)} \in \{1, 2, \ldots, r\},$$

[3]Thomas Bayes, englischer Geistlicher, 1701(?) – 1761

so ergeben sich als Schätzer für die Priors:

$$\widehat{p_j} := \frac{|\{i \in I : y^{(i)} = j\}|}{n}, \tag{8}$$

und für die a-posteriori-Wahrscheinlichkeiten $P(Y = j|X = x)$:

$$\widehat{P}_\Lambda(Y = j|X = x) = \frac{|\{i \in I : x^{(i)} = x \wedge y^{(i)} = j\}|}{|\{i \in I : x^{(i)} = x\}|}. \tag{9}$$

Der sogenannte *Plug-in-Schätzer des Bayes-Klassifikators* b ist dann:

$$\begin{array}{rcl} \widehat{b}_\Lambda : S_X & \to & \{1, \ldots, r\} \\ \widehat{b}_\Lambda(x) & := & k, \text{wobei} \\ \frac{w_k}{\widehat{p}_k}\widehat{P}_\Lambda(Y = k|X = x) & = & \max(\frac{w_j}{\widehat{p}_j}\widehat{P}_\Lambda(Y = j|X = x) : j \in \{1, \ldots, r\}). \end{array} \tag{10}$$

Die Bezeichnung Plug-in-Schätzer rührt daher, dass man in die Formel (7) an allen Stellen Schätzer für die unbekannten Größen einsetzt (to plug in: einstecken).

Beispiel 3.2 (Klassifikation von Pilzen): Wir betrachten die im UCI Machine Learning Repository [UCI] zur Verfügung gestellte Datenmenge »Mushroom« bestehend aus Beschreibungen von 8124 Pilzexemplaren der Gattungen Champignons (lat.: Agaricus) und Schirmlinge (lat.: Lepiota). Zur Beschreibung werden 22 kategorielle Merkmale genutzt, von denen in diesem Beispiel aber nur drei betrachtet werden, nämlich:

- cap-color (Hutfarbe) mit den 10 Ausprägungen: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y;
- gill-color (Lamellenfarbe) mit den 12 Ausprägungen: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y;
- bruises? (Verfärbung bei Druck): true=t, false=f.

Es liegen also drei Merkmale X_1, X_2, X_3 mit endlichen Wertebereichen S_1, S_2, S_3 vor. Der Merkmalsraum S_X ist eine endliche Menge mit $|S_X| = 10 \cdot 12 \cdot 2 = 240$ Elementen. Die Pilze sind in zwei Klassen unterteilt; das Klassenmerkmal Y besitzt die Ausprägungen

Abbildung 8: links: Agaricus campestris, rechts: Lepiota castanea
(Quelle: Wikimedia Commons, ©Andreas Kunze, Lizenz CC BY-SA 3.0)
(Quelle: Wikimedia Commons, ©Ryane Snow, Lizenz CC BY-SA 3.0)

ANMERKUNG: Zur Gattung der Schirmlinge gehört die Pilzart *Lepiota brunneoincarnata*, eine der gefährlichsten Giftpilzarten. Die Spezies dieser Art enthalten u.a. den Giftstoff α-Amanitin, der weder beim Kochen noch im Magen zerstört wird, und nach Aufnahme in den Organismus den Zelltod etwa von Leberzellen bewirkt. Für einen Menschen von 70 Kilogramm Gewicht liegt die tödliche Dosis bei etwa 7 Milligramm.

- p=poisonous (giftig), e=edible (essbar).

Um die Häufigkeitsverteilung der verschiedenen Merkmalsausprägungen von

$$X = (X_1, X_2, X_3) = (\text{cap-color}, \text{gill-color}, \text{bruises?})$$

darzustellen, werden sogenannte *Heat Maps* verwendet, in denen die Häufigkeit farbkodiert erscheint. Da sich die Verteilung dreier kategorieller Merkmale in einem solchen Diagramm nicht darstellen lässt, werden die Merkmalsausprägungen

$$\text{bruises?} = t \text{ und } \text{bruises?} = f$$

separat dargestellt. Um einen Vergleich der Schätzungen (9) der a-posteriori-Wahrscheinlichkeiten zu ermöglichen, werden weiter die Häufigkeitsverteilungen für die beiden Ausprägungen des Klassenmerkmals

$$Y = e \text{ und } Y = p$$

jeweils separat dargestellt.

In Abbildung 9 sind die Verteilungen der Merkmalsausprägungen von cap-color und gill-color bei fester Ausprägung Bruises? $= f$ dargestellt, wobei die Klassen essbarer und giftiger Pilze einander gegenübergestellt werden. Die in den Feldern der Heat Map angegebenen Zahlen sind die absoluten Häufigkeiten. Der Farbton des Feldes gibt von blau nach rot wechselnd steigende Häufigkeiten an. So kann man aus der Abbildung 9 etwa ablesen, dass es in der Stichprobe »Mushroom« 52 essbare Pilze mit den Ausprägungen

$$\text{gill-color} = w,\ \text{cap-color} = w,\ \text{bruises?} = f$$

und 22 giftige Pilze mit den Ausprägungen

$$\text{gill-color} = w,\ \text{cap-color} = n,\ \text{bruises?} = f$$

gibt.

Anmerkung zu den Heat Maps: Die Heat Maps wurden mit der Software Rapidminer 9.7 erstellt. In dieser ist eine durch den Anwender definierte Reihenfolge der Darstellung nominaler Ausprägungen auf den Achsen einer Heat Map nur begrenzt möglich: In den Abbildungen 9 und 10 erscheinen die Ausprägungen des Merkmals gill-color alphabetisch sortiert auf der Ordinate. Es ist dann in Rapidminer leider nicht möglich auch die Ausprägungen von cap-color in alphabetischer Reihenfolge auf der Abszisse darzustellen. Deren Reihenfolge legt Rapidminer nach internen Kriterien fest, wodurch sich diese für zwei Heat Maps unterscheiden kann. In den Abbildungen 9 und 10 tritt diese Situation auf. Den Vergleich der Heat Maps erschwerend kommt noch eine weitere, nicht beeinflussbare Eigenschaft von Rapidminer hinzu: Merkmalausprägungen, zu denen in der in der Heat Map dargestellten Menge keine Samples enthalten sind, werden in der Heat Map weggelassen.

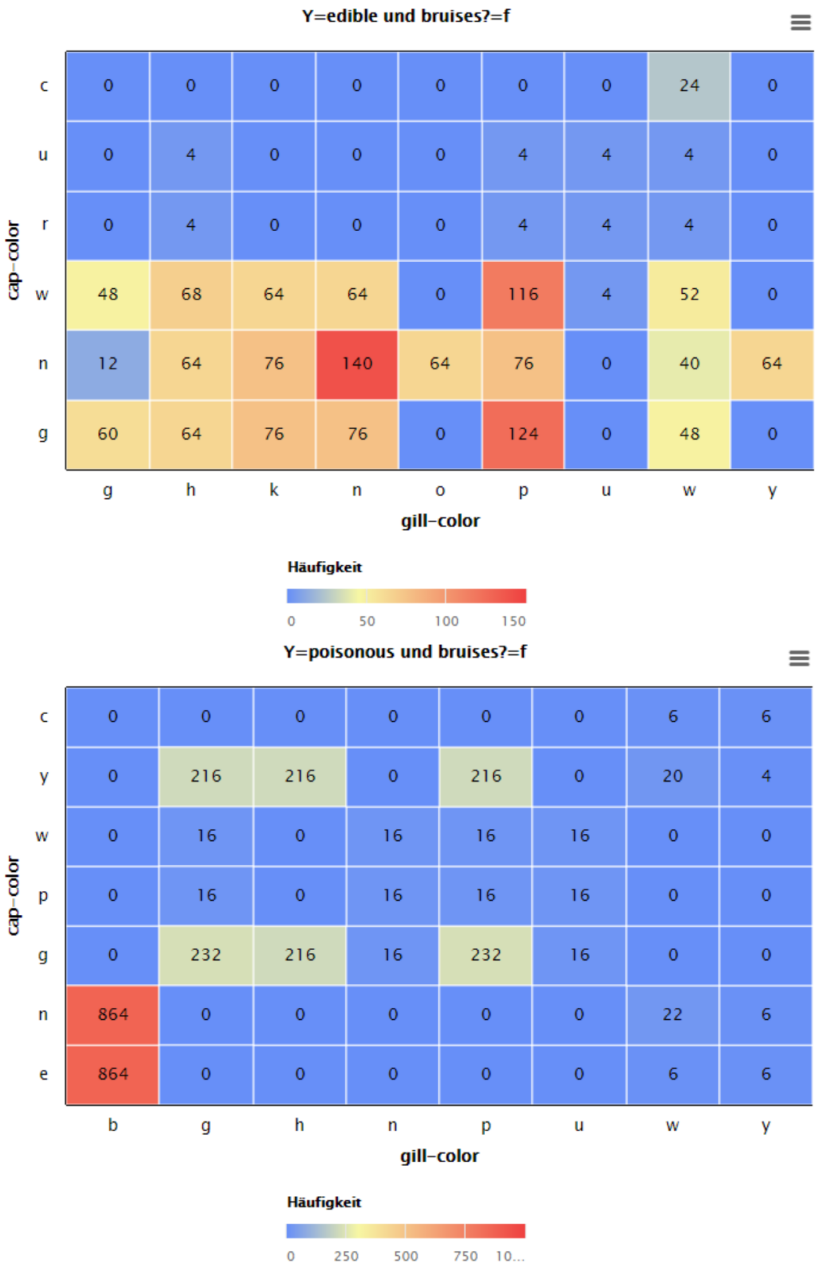


Abbildung 9: Gegenüberstellung der Häufigkeitsverteilung in der Datenmenge »Mushroom« bei bruises?=f: Essbare Pilze oben, Giftpilze unten

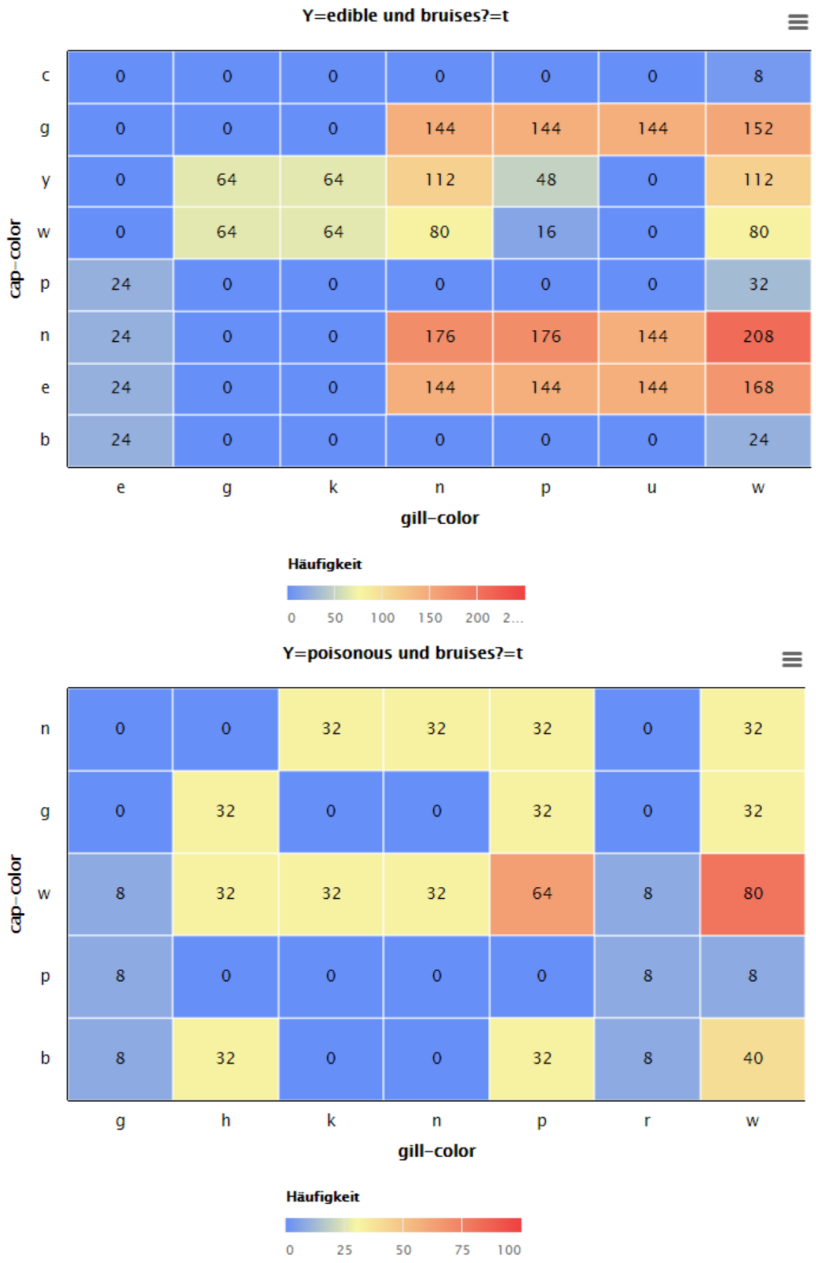


Abbildung 10: Gegenüberstellung der Häufigkeitsverteilung in der Datenmenge »Mushroom« bei Bruises?=t: Essbare Pilze oben, Giftpilze unten

Wir wollen nun einen Pilz mit den Merkmalsausprägungen

$$x := (\text{gill-color} = w,\ \text{cap-color} = n,\ \text{bruises?} = f)$$

im Hinblick auf seine Essbarkeit beziehungsweise Giftigkeit klassifizieren. Dabei soll der aus der Datenmenge »Mushroom« geschätzte Bayes-Klassifikator

$$\widehat{b}_\Lambda : S_X \to \{e, p\}$$

eingesetzt werden; hierbei ist Λ die der Datenmenge unterliegende Stichprobe von Pilzen.

Als Schätzer der Priors (8) ergeben sich:

$$\widehat{p}_{e,\Lambda} = \frac{4208}{8124} = 0.518,\ \widehat{p}_{p,\Lambda} = \frac{3916}{8124} = 0.482.$$

Die Werte wurden auf drei Nachkommastellen gerundet, was auch im Weiteren so gehandhabt wird.

Wegen der hohen Gefährlichkeit mancher giftiger Arten in der Gattung Lepiota gewichtet man die a-posteriori-Wahrscheinlichkeiten $P(Y = p|X = x)$ und $P(Y = e|X = x)$ aus Vorsicht zugunsten der Klassifikation $Y = p$:

$$w_p := 0.7,\ w_e := 0.3.$$

Aus den Heat Maps liest man ab:

$$\widehat{P}_\Lambda(Y = e|X = x) = \frac{40}{40 + 22} = \frac{20}{31}$$

und

$$\widehat{P}_\Lambda(Y = p|X = x) = \frac{22}{40 + 22} = \frac{11}{31}$$

Damit ergibt sich

$$\frac{w_e}{\widehat{p}_{e,\Lambda}} \widehat{P}_\Lambda(Y = e|X = x) = \frac{0.3}{0.518} \cdot \frac{20}{31} \approx 0.374$$

und

$$\frac{w_p}{\widehat{p}_{p,\Lambda}} \widehat{P}_\Lambda(Y = p|X = x) = \frac{0.7}{0.482} \cdot \frac{11}{31} \approx 0.515.$$

Folglich ist

$$\widehat{b}_\Lambda(x) = p,$$

der Pilz wird als giftig eingestuft.

Für einen Pilz mit den Merkmalsausprägungen

$$x := (\text{gill-color} = p,\ \text{cap-color} = g,\ \text{bruises?} = t)$$

erhält man unter Benutzung von Abbildung 10:

$$\widehat{P}_\Lambda(Y = e|X = x) = \frac{144}{144 + 32} = \frac{9}{11}$$

und

$$\widehat{P}_\Lambda(Y = p|X = x) = \frac{32}{144 + 32} = \frac{2}{11}.$$

Es folgt

$$\frac{w_e}{\widehat{p}_{e,\Lambda}} \widehat{P}_\Lambda(Y = e|X = x) = \frac{0.3}{0.518} \cdot \frac{9}{11} \approx 0.474$$

und

$$\frac{w_p}{\widehat{p}_{p,\Lambda}} \widehat{P}_\Lambda(Y = p|X = x) = \frac{0.7}{0.482} \cdot \frac{2}{11} \approx 0.264,$$

womit

$$\widehat{b}_\Lambda(x) = e$$

gilt, der Pilz daher als essbar eingestuft wird.

Die Abbildung 11 ist der Versuch einer Visualisierung der Schätzung $\widehat{b}_\Lambda$ des Bayes-Klassifikators. Hierzu wurden für alle Ausprägungskombinationen $x \in S_X$ die Werte $\widehat{b}_\Lambda(x)$ berechnet, sofern dies möglich ist, das heißt sofern überhaupt Pilze mit der Ausprägungskombination x in der Stichprobe Λ enthalten sind. Die Fälle bruises? $= f$ und bruises? $= t$ sind aus denselben Gründen wie früher wiederum separat dargestellt. Die theoretisch zunächst nicht vorgesehene Klassifikation »unklar« wird dann vergeben, wenn entweder keine Pilze mit der betrachteten Ausprägungskombination in der Stichprobe enthalten sind oder wenn die Gleichung

$$\frac{w_e}{\widehat{p}_{e,\Lambda}} \widehat{P}_\Lambda(Y = e|X = x) = \frac{w_p}{\widehat{p}_{p,\Lambda}} \widehat{P}_\Lambda(Y = p|X = x)$$

gilt.

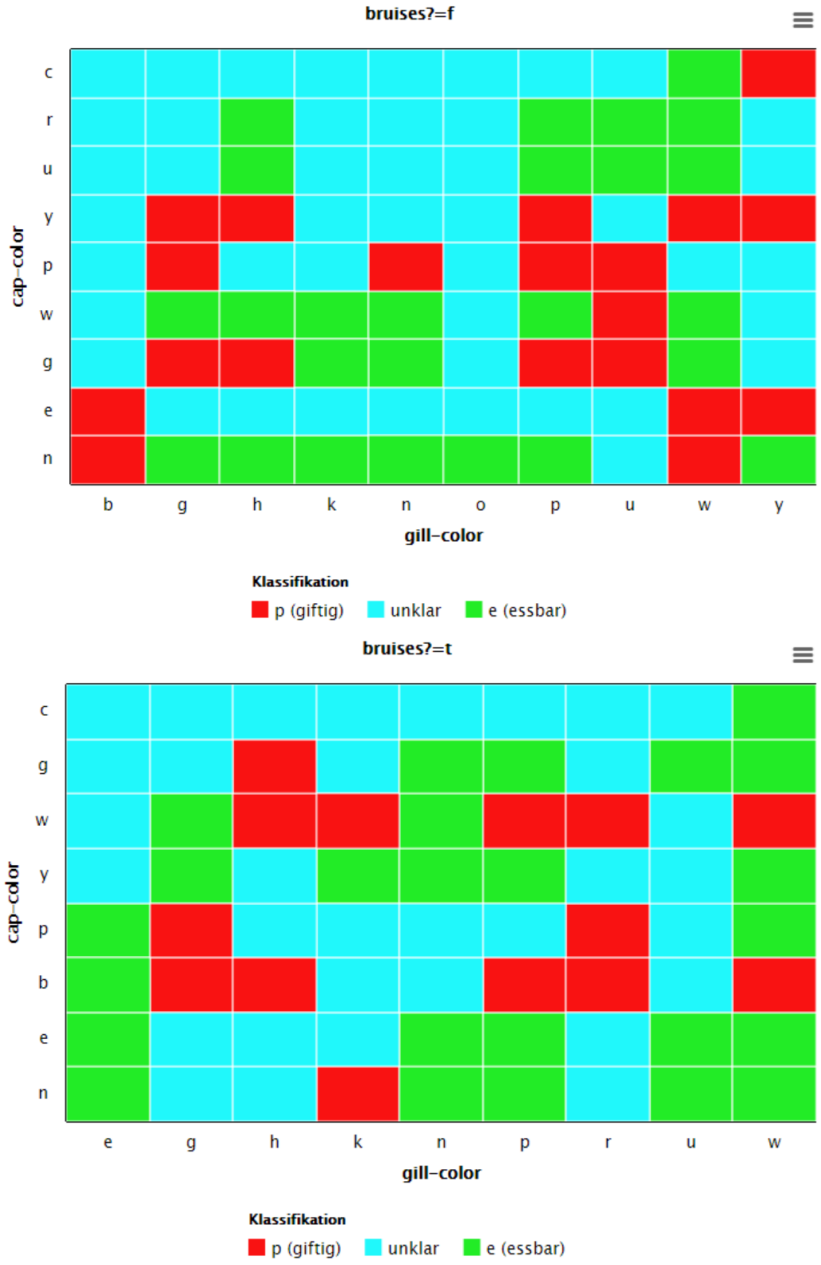


Abbildung 11: Visualisierung der Schätzung $\widehat{b}_\Lambda$ des Bayes-Klassifikators

Für die Darstellung in Abbildung 11 gelten die bereits früher zu Heat Maps gegebenen Hinweise. ◇

Wie das letzte Beispiel zeigt, ist die Durchführung einer Bayes-Klassifikation anhand der vor dem Beispiel gegebenen allgemeinen Ausführungen *im Prinzip* festgelegt. Bei ungünstiger Datenlage enthält die Menge $\{i \in I : x^{(i)} = x\}$ jedoch nur wenige oder gar keine Elemente. In diesem Fall ist eine Schätzung der a-posteriori-Wahrscheinlichkeiten mittels Formel (9) nicht möglich. Im Beispiel tritt dieser Fall etwa für einen Pilz mit den Merkmalsausprägungen

$$x := (\text{gill-color} = g,\ \text{cap-color} = g,\ \text{bruises?} = t)$$

ein. Da es weltweit ungefähr 600 Arten der Gattungen Agaricus und Lepiota gibt und manche Arten seltener als andere sind, kann es sein, dass die hier betrachtete Stichprobe einige dieser Arten nicht repräsentiert. Es gibt (mindestens) zwei Methoden um diese Problematik in der Praxis zu handhaben.

Naive Bayes-Klassifikation

Die zu schätzende a-posteriori-Wahrscheinlichkeit $P(Y = j|X = x)$ lässt sich als

$$P(Y = j|X = x) = p_j \frac{P(X = x|Y = j)}{P_X(x)}$$

schreiben. Man macht nun die

Annahme: *Die Zufallsvariablen $X_1, \ldots, X_p$ sind klassenweise bedingt unabhängig in dem Sinn, dass die Gleichung*

$$P(X = x|Y = j) = \prod_{\ell=1}^{p} P(X_\ell = x_\ell|Y = j) \tag{11}$$

für alle $x \in S_X$ und alle $j \in \{1, \ldots, r\}$ gilt.

Dann erhält man

$$P(Y = j|X = x) = p_j \frac{\prod_{\ell=1}^{p} P(X_\ell = x_\ell|Y = j)}{P_X(x)}.$$

Den nicht von der Klasse j abhängenden Term $P_X(x)$ kann man ignorieren und gelangt zum folgenden

SATZ 3.3: *Es liege das in den Punkten 1 bis 11 von Abschnitt 2.2 formulierte Klassifikationsszenario vor, wobei die Merkmale $X_1, \ldots, X_p$ jeweils abzählbare Ausprägungen besitzen. Weiter sei $\mathbf{F} = \mathrm{M}(S_X, \{1, 2, \ldots, r\})$ und $w_1, \ldots, w_r \in [0, 1]$ seien beliebige Gewichte.*

Unter der Annahme (11) ist dann die Abbildung

$$\begin{aligned} b_n : S_X &\to \{1, \ldots, r\} \\ b_n(x) &:= k, \qquad (12) \\ \frac{w_k}{p_k} \prod_{\ell=1}^{p} P(X_\ell = x_\ell | Y = k) &= \max\left(\frac{w_j}{p_j} \prod_{\ell=1}^{p} P(X_\ell = x_\ell | Y = j) : j \in \{1, \ldots, r\}\right), \end{aligned}$$

eine Lösung des Optimierungsproblems (6).

Man nennt b_n einen naiven Bayes-Klassifikator zu dem gegebenen Klassifikationsproblem.

BEMERKUNG: Der naive Bayes-Klassifikator wird häufig auch dann eingesetzt, wenn die Annahme (11) nicht erfüllt ist und kann dennoch gute Ergebnisse liefern.

Eine Schätzung $\widehat{b}_{n,\Lambda}$ von b_n anhand einer Datenmenge

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}),\ x^{(i)} \in S_X, y^{(i)} \in \{1, 2, \ldots, r\}$$

zu einer Stichprobe $\Lambda \subseteq \Omega$ ermittelt man nun wie folgt: Die apriori-Wahrscheinlichkeiten p_j werden wie früher gemäß

$$\widehat{p}_{j,\Lambda} := \frac{|\{i \in I : y^{(i)} = j\}|}{n}$$

geschätzt. Schätzer für die merkmalsweisen, bedingten Wahrscheinlichkeiten liefert

$$\widehat{P}_\Lambda(X_\ell = x_\ell | Y = j) := \frac{|\{i \in I : x_\ell^{(i)} = x_\ell \wedge y^{(i)} = j\}|}{|\{i \in I :\ y^{(i)} = j\}|}.$$

Der Plug-in-Schätzer für den naiven Bayes-Klassifikators b_n ist dann:

$$\begin{aligned} \widehat{b}_{n,\Lambda} : S_X &\to \{1, \ldots, r\} \\ \widehat{b}_{n,\Lambda}(x) &:= k, \text{wobei} \qquad (13) \\ \frac{w_k}{\widehat{p}_{k,\Lambda}} \prod_{\ell=1}^{p} \widehat{P}_\Lambda(X_\ell = x_\ell | Y = k) &= \max\left(\frac{w_j}{\widehat{p}_{j,\Lambda}} \prod_{\ell=1}^{p} \widehat{P}_\Lambda(X_\ell = x_\ell | Y = j) : j \in \{1, \ldots, r\}\right). \end{aligned}$$

Der Vorteil des naiven Bayes-Klassifikators im Vergleich zum gewöhnlichen ist, dass weniger Größen zu schätzen sind und die beim gewöhnlichen Bayes-Klassifikator dabei möglicherweise auftretenden Problem im naiven Fall nicht auftreten können: Verschwindende Nenner können nur im Fall einer leeren Stichprobenklasse $\Lambda_j = \Omega_j \cap \Lambda$ auftreten und in einem solchen Fall ist eine Erweiterung der Stichprobe unvermeidbar.

Beispiel 3.4 (Klassifikation von Pilzen (Forts.)): Zur Schätzung eines naiven Bayes-Klassifikators im Beispiel 3.2 müssen insgesamt

$$2 \cdot (10 + 12 + 2) + 2 = 50$$

Werte aus der Datenmenge berechnet werden, im Unterschied zu den

$$2 \cdot (10 \cdot 12 \cdot 2) + 2 = 482$$

Werten im Fall des gewöhnlichen Bayes-Klassifikators. Dies sind zunächst die Schätzungen für die beiden a-priori-Wahrscheinlichkeiten:

$$\widehat{p}_{e,\Lambda} = \frac{4208}{8124} = 0.518,\ \widehat{p}_{p,\Lambda} = \frac{3916}{8124} = 0.482.$$

Für jede Ausprägung $x_i \in S_{X_i}$ jedes der drei Merkmale X_i sind dann die Werte $\widehat{P}_\Lambda(X_i = x_i|Y = e)$ und $\widehat{P}_\Lambda(X_i = x_i|Y = p)$ zu berechnen. Die Häufigkeitsverteilungen von $X_1 =$ cap-color und $X_2 =$ gill-color sind gruppiert nach den Ausprägungen von Y in Abbildung 12 dargestellt. Für $X_3 =$ bruises? ergibt sich:

$$\widehat{P}_\Lambda(X_3 = t|Y = e) = \frac{2752}{4208} = 0.654, \quad \widehat{P}_\Lambda(X_3 = t|Y = p) = \tfrac{624}{3916} = 0.159,$$
$$\widehat{P}_\Lambda(X_3 = f|Y = e) = \frac{1456}{4208} = 0.346, \quad \widehat{P}_\Lambda(X_3 = f|Y = p) = \tfrac{3292}{3916} = 0.841.$$

Will man nur einen konkreten Pilz klassifizieren, reduziert sich der Aufwand auf diejenigen Ausprägungen, die der Pilz besitzt.

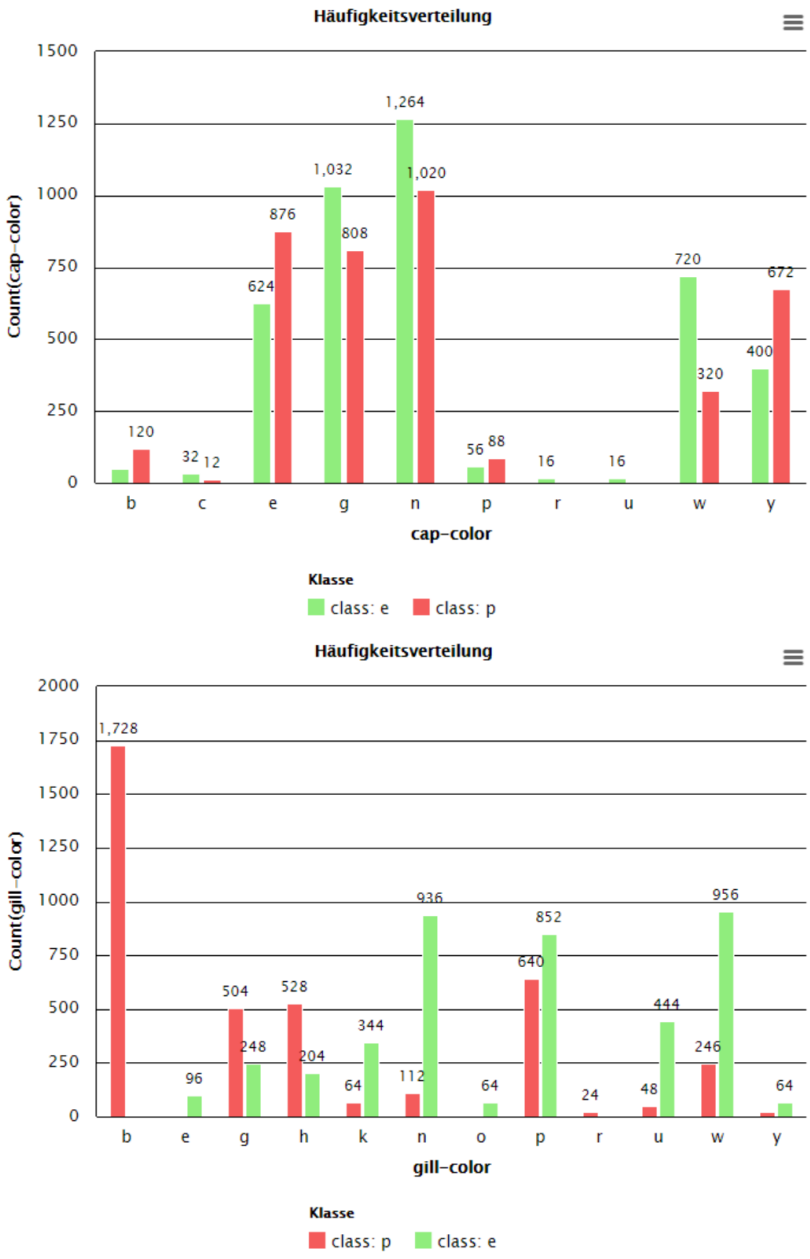


Abbildung 12: Häufigkeitsverteilungen von cap-color und gill-color

Wir betrachten einen der beiden bereits im Beispiel 3.2 klassifizierten Pilze. Dieser besitzt die Ausprägungen

$$x := (\text{gill-color} = w,\ \text{cap-color} = n,\ \text{bruises?} = f).$$

Hierzu gilt bei Rundung auf drei Nachkommastellen:

$$\begin{aligned}
\widehat{P}_\Lambda(\text{gill-color} = w|Y = e) &= \tfrac{956}{4208} = 0.227\\
\widehat{P}_\Lambda(\text{gill-color} = w|Y = p) &= \tfrac{246}{3916} = 0.063\\
\widehat{P}_\Lambda(\text{cap-color} = n|Y = e) &= \tfrac{1264}{4208} = 0.300\\
\widehat{P}_\Lambda(\text{cap-color} = n|Y = p) &= \tfrac{1020}{3916} = 0.260\\
\widehat{P}_\Lambda(\text{bruises?} = f|Y = e) &= \tfrac{1456}{4208} = 0.346\\
\widehat{P}_\Lambda(\text{bruises?} = f|Y = p) &= \tfrac{3292}{3916} = 0,841.
\end{aligned}$$

Benutzt man wieder die Gewichte $w_p = 0.7$ und $w_e = 0.3$, so ergeben sich für die Größen

$$\begin{aligned}
t_p &:= \tfrac{w_p}{\widehat{p}_{p,\Lambda}}\widehat{P}_\Lambda(\text{gill-color} = w|Y = p)\widehat{P}_\Lambda(\text{cap-color} = n|Y = p)\widehat{P}_\Lambda(\text{bruises?} = f|Y = p)\\
t_e &:= \tfrac{w_e}{\widehat{p}_{e,\Lambda}}\widehat{P}_\Lambda(\text{gill-color} = w|Y = e)\widehat{P}_\Lambda(\text{cap-color} = n|Y = e)\widehat{P}_\Lambda(\text{bruises?} = f|Y = e)
\end{aligned}$$

die Werte

$$\begin{aligned}
t_p &= \tfrac{0.7}{0.482} \cdot 0.063 \cdot 0.260 \cdot 0,841 = 0.020\\
t_e &= \tfrac{0.3}{0.518} \cdot 0.227 \cdot 0.300 \cdot 0.346 = 0,015.
\end{aligned}$$

Alle Werte wurden wieder auf drei Nachkommastellen gerundet. Es folgt

$$\widehat{b}_{n,\Lambda}(x) = p,$$

der Pilz wird also wie im Fall des gewöhnlichen Bayes-Klassifikators als giftig klassifiziert. ◇

Nächste-Nachbarn-Klassifikation

Eine Alternative zum naien Bayes-Klassifikator im Fall leerer oder zu kleiner Mengen $\{i \in I : x^{(i)} = x\}$ ist der in Abschnitt 2.1 eingeführte Nächste-Nachbarn-Klassifikator: Anstatt diejenigen Samples in der Datenmenge zu einer Stichprobe Λ zu betrachten, für die $x^{(i)} = x$ gilt, erweitert man den Blickwinkel etwas und fasst auch diejenigen Samples ins Auge, für die der Abstand $d(x^{(i)}, x)$ bezüglich einer zum Anwendungskontext passenden Metrik $d : S_X \times S_X \to \mathbb{R}^{\geq 0}$ hinreichend klein ist bzw. betrachtet die bezüglich der Metrik d nächsten Nachbarn von x.

Beispiel 3.5: Wir betrachten erneut die im Abschnitt 2.1 eingeführte Datenmenge »Olive Oils« und beschränken uns auf die Merkmale »Ölsäure« und »Linolsäure«. In dem dort beschriebenen Experiment zur Güte von k-Nächste-Nachbarn-Klassifikatoren hat sich gezeigt, dass die Wahl von $k = 3$ oder $k = 5$ angemessen sein könnte – siehe die Abbildung 7.

Um eine Visualisierung des 5-NN-Klassifikators $f_{5,\Lambda}$ zu der Stichprobe Λ, auf der die Datenmenge »Olive Oils« basiert, zu erhalten, wurden 10000 in der Teilmenge

$$[6300, 8410] \times [448, 1470] \subset S_X$$

zufällig uniform verteilte Samples erzeugt und mit $f_{5,\Lambda}$ klassifiziert. Die Klassifikationsergebnisse sind in Abbildung 13 farbkodiert dargestellt.

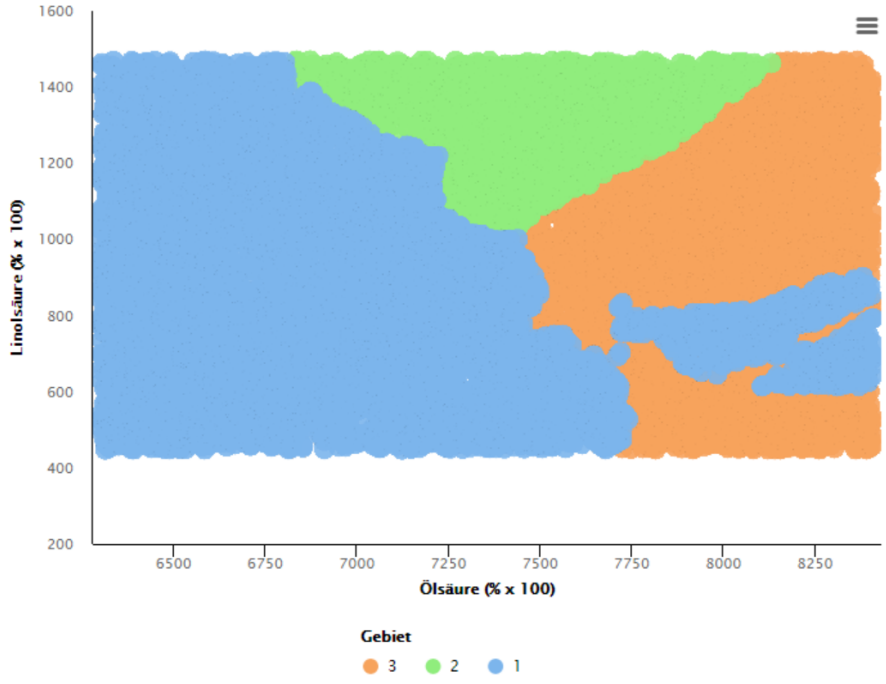


Abbildung 13: Klassifikationsgebiete $f_{5,\Lambda}^{-1}(k)$, $k \in \{1, 2, 3\}$, auf einer Teilmenge des Merkmalsraums von »Olive Oils«: 1: Süditalien, 2: Sardinien, 3: Norditalien

Anmerkung: Die Farbfüllung der Klassifikationsgebiete in Abbildung 13 wurde durch die Wahl großer Datenpunktmarker in Rapidminer erreicht. ◇

3.2 Reellwertige Inputs

Wir machen in diesem Abschnitt die Annahme, dass die Zufallsvariablen X_k sämtlich reellwertig und die Wertebereiche $S_k \subseteq \mathbb{R}$ Intervalle einer beliebigen Art sind. Als σ-Algebra auf S_k wählt man jeweils die von der Borelschen σ-Algebra $\mathcal{B}$ induzierte:

$$\mathcal{S}_k := \{B \cap S_k : B \in \mathcal{B}\}.$$

Damit ergibt sich, dass die Produkt-σ-Algebra

$$\mathcal{S} = \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_p$$

auf dem Merkmalsraum $S_X = S_1 \times \ldots \times S_p$ von der Borelschen σ-Algebra

$$\mathcal{B}^p = \mathcal{B} \otimes \ldots \otimes \mathcal{B}$$

induziert wird.

Der Merkmalsraum S_X ist als Produkt von Intervallen selbst eine Borelmenge: $S_X \in \mathcal{B}^p$. In dieser Situation kann man sich Wahrscheinlichkeitsmaße auf $\mathcal{S}$ über Lebesgue-integrierbare Funktionen verschaffen:

Satz 3.6 ([Geo], Satz 1.18): *Es sei $D \in \mathcal{B}^p$ und $\varrho : D \to [0, \infty)$ eine Funktion mit den Eigenschaften:*

- $\forall c \in \mathbb{R}^{>0} \quad \{x \in D : \varrho(x) \leq c\} \in \mathcal{B}^p$,
- $\int\limits_D \varrho(x)\, dx = 1.$

Dann ist auf der durch $\mathcal{B}^p$ auf D induzierten σ-Algebra $\mathcal{D}$ ein Wahrscheinlichkeitsmaß durch

$$P(E) := \int\limits_E \varrho(x)\, dx$$

gegeben. Man bezeichnet ϱ als Dichtefunktion zu P.

Für jede der r Klassen Ω_k kann man die Zufallsvariable

$$Z_k := X|_{\Omega_k} : \Omega_k \to S_X$$

und das zugehörige Bildmaß

$$P_{Z_k} : \mathcal{S} \to [0,1],\ A \mapsto P_k(Z_k^{-1}(A))$$

betrachten.

Wir machen nun für das Weitere die Annahme: *Die Bildmaße P_{Z_k} sind alle gemäß Satz 3.6 durch Wahrscheinlichkeitsdichten ϱ_k definiert.*

Es sei $f : S_X \to \{1, 2, \ldots, r\}$ ein messbarer Klassifikator und $w_1, \ldots, w_r \in [0,1]$ seien wiederum Gewichte für die einzelnen Klassen. Dann gilt für die gewichtete, totale Trefferwahrscheinlichkeit von f:

$$\begin{aligned}
P(f(X_1, \ldots, X_p) = Y)_w &= \sum_{k=1}^{r} w_k P(f(X) = k | Y = k) \\
&= \sum_{k=1}^{r} w_k P(X \in f^{-1}(k) | Y = k) \\
&= \sum_{k=1}^{r} w_k P_k(Z_k \in f^{-1}(k)) \\
&= \sum_{k=1}^{r} w_k P_{Z_k}(f^{-1}(k)) \\
&= \sum_{k=1}^{r} w_k \int\limits_{f^{-1}(k)} \varrho_k(x)\, dx \\
&= \sum_{k=1}^{r} \int\limits_{f^{-1}(k)} w_k \varrho_k(x)\, dx.
\end{aligned}$$

Die gewichtete, totale Trefferwahrscheinlichkeit wird also genau dann maximal, wenn jeder der nicht-negativen Summanden $\int\limits_{f^{-1}(k)} w_k \varrho_k(x)\, dx$ maximal wird, wenn f die Menge S_X also so in die disjunkten Teilmengen $f^{-1}(k)$, $k \in \{1, \ldots, r\}$, zerlegt, dass jedes $x \in S_X$ in derjenigen Teilmenge $f^{-1}(k)$ liegt, für die

$$w_k \varrho_k(x) = \max(w_1 \varrho_1(x), \ldots, w_r \varrho_r(x))$$

gilt. Wir haben bewiesen:

Satz 3.7: *Es liege das in den Punkten 1 bis 11 von Abschnitt 2.2 formulierte Klassifikationsszenario vor, wobei die Merkmale $X_1, \ldots, X_p$ jeweils reelle Intervalle $S_k \subseteq \mathbb{R}$ als Wertebereiche besitzen.*

Weiter seien die Bildmaße P_{Z_k} zu den Zufallsvariablen $Z_k := X|_{\Omega_k}$ alle gemäß Satz 3.6 durch Wahrscheinlichkeitsdichten ϱ_k definiert.

Schließlich sei $\mathbf{F} = \mathrm{M}(S_X, \{1, 2, \ldots, r\})$ und $w_1, \ldots, w_r \in [0, 1]$ seien beliebige Gewichte.

Dann ist die Abbildung

$$\begin{array}{rcl} b : S_X & \to & \{1, \ldots, r\} \\ b(x) & := & k, \\ \textit{wobei } w_k\varrho_k(x) & = & \max(w_j\varrho_j(x) : j \in \{1, \ldots, r\}) \end{array} \tag{14}$$

eine Lösung des Optimierungsproblems (6).

Man nennt b einen Bayes-Klassifikator zu dem gegebenen Klassifikationsproblem.

Bemerkung: Die Eigenschaft, dass P_{Z_k} durch die Dichte ϱ_k gegeben ist, drückt man kurz auch so aus: Die Zufallsvariable Z_k ist ϱ_k-verteilt.

Normalverteilte Zufallsvariablen Z_k

Ein in der Anwendung besonders verbreiteter Spezialfall von Satz 3.7 ist durch folgende Rahmenbedingungen gegeben:

- Für alle k ist $S_k = \mathbb{R}$.

  Dann ist also $S_X = \mathbb{R}^p$ und $\mathcal{S} = \mathcal{B}^p$.

- Die Zufallsvariablen Z_k sind normalverteilt:

$$\varrho_k(x) = N(\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)}, \tag{15}$$

  wobei $\mu_k \in \mathbb{R}^p$ der *Erwartungswert* und $\Sigma_k \in \mathbb{R}^{p\times p}$ die *Kovarianzmatrix* von Z_k sind.

Da der natürliche Logarithmus eine streng monoton wachsende Funktion ist, gilt die Maximums-Bedingung in (14) genau dann, wenn die Bedingung

$$\ln(w_k\varrho_k(x)) = \max(\ln(w_j\varrho_j(x)) : j = 1, \ldots, r) \tag{16}$$

gilt. Nun ist aber

$$\ln(w_j\varrho_j(x)) = -\frac{1}{2}(x-\mu_j)^t\Sigma_j^{-1}(x-\mu_j) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(\Sigma_j)) + \ln(w_j),$$

womit insgesamt dasjenige $k \in \{1, \ldots, r\}$ zu ermitteln ist, für welches

$$(x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) + \ln(\det(\Sigma_k)) - 2\ln(w_k)$$

gleich dem Minimum

$$\min((x - \mu_j)^t \Sigma_j^{-1} (x - \mu_j) + \ln(\det(\Sigma_j)) - 2\ln(w_j) : j = 1, \ldots, r) \qquad (17)$$

ist. Man beachte, dass die Terme $(x - \mu_j)^t \Sigma_j^{-1} (x - \mu_j)$ Polynome zweiten Grades in den p Variablen $x_1, \ldots, x_p$ sind.

Wir diskutieren verschiedene Fälle zur Illustration ausführlicher.

Der Fall $p = 1$ und $r = 2$

Es liegt nur ein Merkmal $X = X_1$ vor. In diesem Fall gilt $\Sigma_j = \sigma_j^2$, wobei σ_j^2 die Varianz der Zufallsvariablen $Z_j = X|_{\Omega_j}$ ist. Für den Bayes-Klassifikator gilt also $b(x) = k$, falls

$$\frac{(x - \mu_k)^2}{\sigma_k^2} + \ln(\sigma_k^2) - 2\ln(w_k) = \min(\frac{(x - \mu_j)^2}{\sigma_j^2} + \ln(\sigma_j^2) - 2\ln(w_j) : j = 1, \ldots, r).$$

Im Fall $r = 2$ von nur zwei Klassen kann man dies auch folgendermaßen schreiben:

$$b(x) := \begin{cases} 1 & \text{falls } \frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2} + \ln(\frac{\sigma_1^2}{\sigma_2^2}) - 2\ln(\frac{w_1}{w_2}) < 0; \\ 2 & \text{falls } \frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2} + \ln(\frac{\sigma_1^2}{\sigma_2^2}) - 2\ln(\frac{w_1}{w_2}) > 0. \end{cases} \qquad (18)$$

Im Fall

$$\frac{(x - \mu_1)^2}{\sigma_1^2} - \frac{(x - \mu_2)^2}{\sigma_2^2} + \ln(\frac{\sigma_1^2}{\sigma_2^2}) - 2\ln(\frac{w_1}{w_2}) = 0 \qquad (19)$$

kann man x streng genommen nicht klassifizieren. Man kann jedoch je nach Kontext x auch willkürlich einer der beiden Klassen zuordnen.

Die quadratische Gleichung (19) kann zwei, eine oder keine Lösungen besitzen.

Zwei Lösungen: Sind $a, b \in \mathbb{R}$, $a < b$, die Lösungen von (19), so besitzt die Bayes-Klassifikationsregel eine der beiden folgenden Formen:

- $b(x) = 1 \Leftrightarrow x \in (a, b)$,
- $b(x) = 2 \Leftrightarrow x \in (a, b)$.

Welcher der beiden Fälle eintritt hängt von den Werten der Parameter w_1, w_2, μ_1, μ_2, σ_1 und σ_2 ab. Betrachtet man zum Beispiel den Fall

$$w_1 = 0.8, \sigma_1 = 0.2, \mu_1 = 0, w_2 = 0.2, \sigma_2 = 0.005, \mu_2 = 0.5,$$

so zeigt die Abbildung 14 die entstehende Situation. Es gilt:

$$a \approx 0.41,\ b \approx 0.62$$

und die Klassifikationsregel

$$b(x) = 2 \Leftrightarrow x \in (a, b).$$

Die in der Abbildung dargestellte Mischdichte $\varrho := w_1\varrho_1 + w_2\varrho_2$ liefert die Verteilung der Zufallsvariablen $w_1Z_1 + w_2Z_2$. Diese ist gleich der Verteilung von X selbst, falls $w_1 = p_1$ und $w_2 = p_2$ gewählt wird – siehe Punkt 5 des allgemeinen Klassifikationsszenarios.

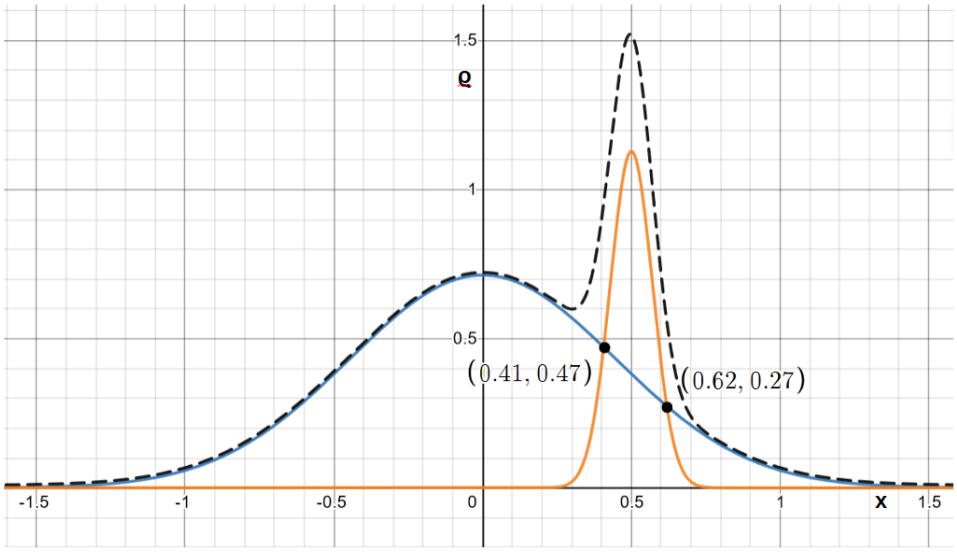


Abbildung 14: 2-Klassen-Bayes-Klassifikation bei einem reellen Merkmal X
blaue Kurve: $w_1\varrho_1 = 0.8N(0, 0.2)$, orange Kurve: $w_2\varrho_2 = 0.2N(0.5, 0.005)$,
schwarz-durchbrochene Kurve: Mischdichte ϱ.

Vergleichbare Dichten $w_1\varrho_1$ und $w_2\varrho_2$: Gilt eine der beiden Ungleichungen

$$\forall x \in \mathbb{R} \quad w_1\varrho_1(x) \leq w_2\varrho_2(x)$$

oder

$$\forall x \in \mathbb{R} \quad w_1\varrho_1(x) \geq w_2\varrho_2(x),$$

so besitzt die Gleichung (19) keine oder genau eine Lösung. Die Bayes-Klassifikationsregel klassifiziert dann abhängig davon, welche der Ungleichungen gilt, *jedes* x in die Klasse 1 oder *jedes* x in die Klasse 2. Zur Unterscheidung der Klassen ist die Regel also wertlos. Diese Situation tritt in der Praxis zum Beispiel dann ein, wenn die Zugehörigkeit zu einer von beiden Klassen eine sehr kleine a-priori-Wahrscheinlichkeit besitzt, also etwa in der medizinischen Diagnose einer seltenen Krankheit.

In Abbildung 15 ist der konkrete Fall

$$w_1 = 0.95, \sigma_1 = 0.2, \mu_1 = 0, w_2 = 0.05, \sigma_2 = 0.005, \mu_2 = 0.5,$$

dargestellt, in dem alle Samples als Klasse 1 klassifiziert werden. Man beachte, dass als Hinweis auf das Vorhandensein der Klasse 2 in der Mischdichte ϱ durchaus eine zweites Maximum auftritt, das aber von der Bayes-Regel nicht verwendet wird.

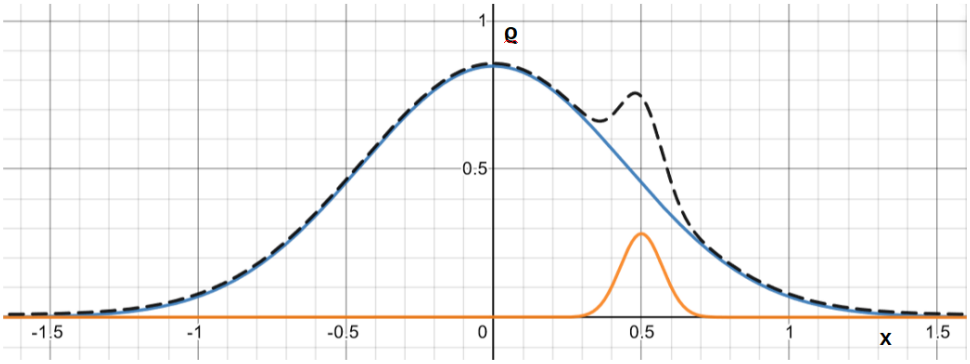


Abbildung 15: 2-Klassen-Bayes-Klassifikation bei einem reellen Merkmal X
blaue Kurve: $w_1\varrho_1 = 0.95N(0, 0.2)$, orange Kurve: $w_2\varrho_2 = 0.05N(0.5, 0.005)$,
schwarz-durchbrochene Kurve: Mischdichte ϱ.

Eine Lösung bei nicht vergleichbaren Dichten: Nimmt man an, dass die gewichteten Dichtefunktionen $w_1\varrho_1$ und $w_2\varrho_2$ nicht wie im vorangegangenen Fall vergleichbar sind und dass die Gleichung (19) genau eine Lösung a besitzt, so nimmt die Bayes-Klassifikationsregel eine der beiden folgenden Formen an:

- $b(x) = 1 \Leftrightarrow x < a$,
- $b(x) = 2 \Leftrightarrow x < a$,.

Im Fall $x = a$ kann streng genommen wieder keine Klassifikationsaussage gemacht werden.

Welcher der beiden Fälle eintritt hängt von den Werten der Parameter w_1, w_2, μ_1, μ_2, σ_1 und σ_2 ab. Die Abbildung 16 stellt den Fall

$$w_1 = 0.6, \sigma_1 = 0.1, \mu_1 = 0, w_2 = 0.4, \sigma_2 = 0.1, \mu_2 = 0.5,$$

mit $a \approx 0.33$ dar.

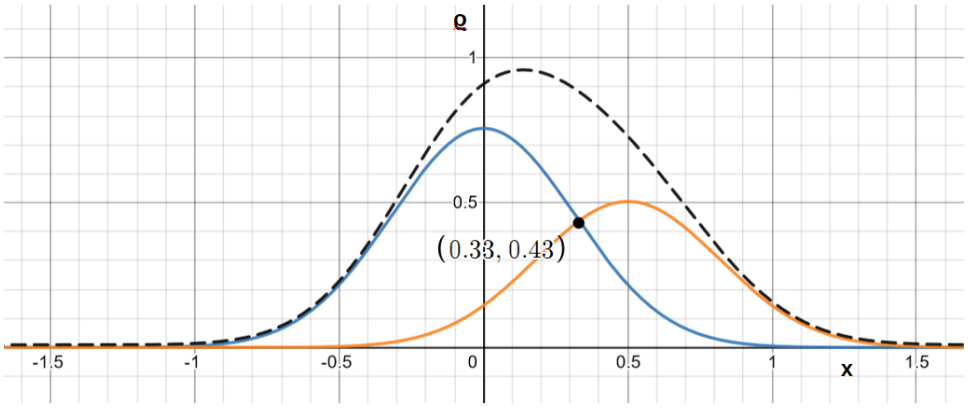


Abbildung 16: 2-Klassen-Bayes-Klassifikation bei einem reellen Merkmal X blaue Kurve: $w_1\varrho_1 = 0.6N(0, 0.1)$, orange Kurve: $w_2\varrho_2 = 0.4N(0.5, 0.1)$, schwarz-durchbrochene Kurve: Mischdichte ϱ.

Der Fall $p = 2$ und $r = 2$

In diesem Fall kann man den Bayes-Klassifikator analog zu (18) mit Hilfe der Testgröße

$$T(x) := (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)-(x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2)+\ln(\frac{\det(\Sigma_1)}{\det(\Sigma_2)})-2\ln(\frac{w_1}{w_2}) \quad (20)$$

in der Form

$$b(x) := \begin{cases} 1 & \text{falls} \quad T(x) < 0 \\ 2 & \text{falls} \quad T(x) > 0 \end{cases} \quad (21)$$

schreiben, wobei $\mu_1, \mu_2 \in \mathbb{R}^2$ und $\Sigma_1, \Sigma_2 \in \mathbb{R}^{2\times 2}$. Die Lösungsmenge der Gleichung

$$(x-\mu_1)^t)\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) + \ln(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}) - 2\ln(\frac{w_1}{w_2}) = 0$$

trennt also die beiden Klassen. Es handelt sich um eine quadratische Gleichung in zwei Variablen, nämlich den Komponenten x_1, x_2 des Merkmalsvektors $x \in \mathbb{R}^2$. Folglich ist die Lösungsmenge ein Kegelschnitt – siehe Abbildung 17.

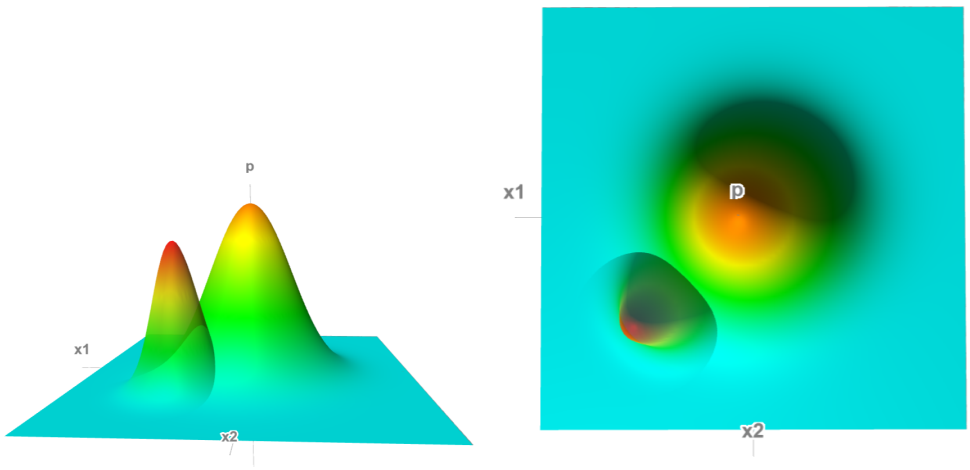


Abbildung 17: Schnitt der Graphen zweier Gauß-Dichten
(Beleuchtung von links vorne)

Wir kehren zurück zur Betrachtung des Klassifikators (14) bei klassenweiser Normalverteilung aber bei beliebigem p. Um diesen aus den Daten $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ zu einer Stichprobe $\Lambda \subseteq \Omega$ zu schätzen, müssen die Erwartungswerte μ_k und die Kovarianzmatrizen Σ_k in jeder Klasse geschätzt werden. Hierzu verwendet man die Schätzer:

- $\widehat{\mu}_k = \frac{1}{n_k} \sum\limits_{y^{(i)}=k} x^{(i)}$, wobei $n_k := |\{i \in I : y^{(i)} = k\}|$,
- $\widehat{\Sigma}_k := \frac{1}{n_k} \sum\limits_{y^{(i)}=k} (x^{(i)} - \widehat{\mu}_k)(x^{(i)} - \widehat{\mu}_k)^t$.

Liegen zwischen den einzelnen Merkmalen keine oder nur schwache Korrelationen vor, so kann man annehmen, dass die Kovarianzmatrizen Σ_k Diagonalmatrizen sind. In diesem Fall ergeben sich die Diagonalkoeffizienten von $\widehat{\Sigma}_k$ durch Schätzung der klassenweisen Varianzen der X_j:

$$\widehat{\sigma}^2_{j,k} := \frac{1}{n_k - 1} \sum_{y^{(i)}=k} (x^{(i)}_j - \widehat{\mu}_{k,j})^2.$$

Da die Diagonalitätsannahme die Anzahl zu schätzender Parameter von $\frac{1}{2}p^2 + p$ auf p reduziert, ist diese inbesondere bei kleinen Stichproben Λ in Erwägung zu ziehen.

Beispiel 3.8: Wir betrachten die Datenmenge »Seeds« aus dem UCI-Machine-Learning-Repository [UCI]: Bei den 210 Samples dieser Datenmenge handelt es sich um Messungen geometrischer Parameter der Körner von drei verschiedenen Weizenarten, nämlich den Arten »Kama« (Klasse 1), »Rosa« (Klasse 2) und »Canadian« (Klasse 3). Im vorliegenden Beispiel werden zwei der sieben Merkmalen der Datenmenge verwendet, nämlich

$$\begin{array}{rcl} X_1 & := & \text{A (Querschnittsfläche des Korns),} \\ X_2 & := & \text{AC (Asymmetrie des Korns)} \end{array}$$

– siehe Abbildung 18. Die Einheiten sind in der Datenbeschreibung nicht angegeben, sind aber wohl Quadratmillimeter (mm^2) für A, während AC dimensionslos ist. Das Klassenmerkmal ist

$$Y : \Omega \to \{1, 2, 3\};$$

die einzelnen Klassen sind in der Stichprobe jeweils mit 70 Samples vertreten.

Abbildung 18: Weizenkörner

Für die klassenweisen Mittelwerte ergibt sich:

$$\widehat{\mu}_1 = (14.334, 2.667),\ \widehat{\mu}_2 = (18.334, 3.645),\ \widehat{\mu}_3 = (11.874, 4.788).$$

Die geschätzten Kovarianzmatrizen sind:

$$\widehat{\Sigma}_1 = \begin{pmatrix} 1.478 & -0.072 \\ -0.072 & 1.378 \end{pmatrix},\ \widehat{\Sigma}_2 = \begin{pmatrix} 2.072 & -0.067 \\ -0.067 & 1.397 \end{pmatrix},\ \widehat{\Sigma}_3 = \begin{pmatrix} 0.523 & 0.038 \\ 0.038 & 1.786 \end{pmatrix}.$$

Weiter gilt

$$\widehat{p}_1 = \widehat{p}_2 = \widehat{p}_3 = \frac{1}{3}.$$

Verwendet man die a-priori-Wahrscheinlichkeiten als Gewichte, so liegen nun Schätzer aller im Bayes-Klassifikator (17) vorkommender Größen vor. Die resultierende Schätzung dieses Klassifikators ist in Abbildung 19 dargestellt, wobei die Darstellungsform der im Beispiel 3.5 verwendeten entspricht. Für die Darstellung wurden 9000 uniform verteilte, zufällige Samples verwendet. ◇

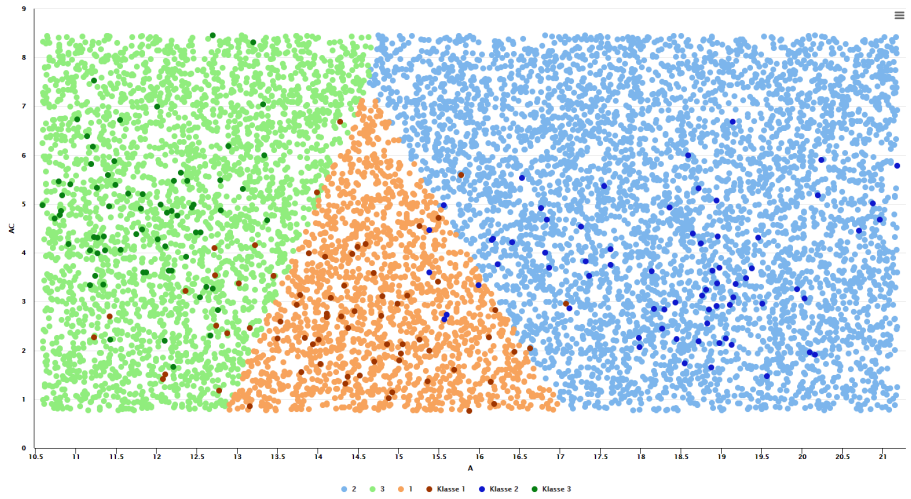


Abbildung 19: Klassifikationsgebiete eines basierend auf der Datenmenge »Seeds« unter der Annahme klassenweiser Normalverteilung erstellten Bayes-Klassifikators. Die Datenmenge »Seeds« ist in dunkleren Farbtönen dargestellt.

4 Güteschätzung für Klassifikatoren

Wir betrachten in diesem Abschnitt erneut ein allgemeines Klassifikationsproblem aus stochastischer Sicht, wie im Abschnitt 2.2 beschrieben. Wie dort sei $\Lambda \subseteq \Omega$ eine Stichprobe und

$$\widehat{f}_\Lambda : S_X \to \{1, 2, \ldots, r\}$$

ein Schätzer basierend auf der Stichprobe Λ für eine Lösung des Optimierungsproblems (6) in einem gegebenen Suchraum $\mathbf{F} \subseteq \mathrm{M}(S_X, \{1, 2, \ldots, r\})$ und bezüglich gegebener Gewichte $w = (w_1, w_2, \ldots, w_r)$. Das naheliegendste Maß für die Güte von $\widehat{f}_\Lambda$ ist dann die gewichtete, totale Trefferwahrscheinlichkeit (Definition 2.1)

$$P(\widehat{f}_\Lambda(X_1, \ldots, X_p) = Y)_w = \sum_{k=1}^{r} w_k P(\widehat{f}_\Lambda(X_1, \ldots, X_p) = k | Y = k).$$

Diese kann allerdings selbst nur anhand der Stichprobe Λ geschätzt werden, da das Wahrscheinlichkeitsmaß P unbekannt ist.

Man betrachtet die Zufallsvariable

$$H : \Omega \to \{0,1\}, \omega \mapsto \begin{cases} 0 & \text{falls } \widehat{f}_\Lambda(X_1(\omega), \ldots, X_p(\omega)) \neq Y(\omega) \\ 1 & \text{falls } \widehat{f}_\Lambda(X_1(\omega), \ldots, X_p(\omega)) = Y(\omega). \end{cases}$$

Für deren Erwartungswert gilt

$$E(H) = P(\widehat{f}_\Lambda(X_1, \ldots, X_p) = Y),$$

womit die totale Trefferquote (Definition 2.1) $T(\widehat{f}_\Lambda, \Lambda)$ ein Schätzer für die totale Trefferwahrscheinlichkeit ist. Analog ist

$$T(\widehat{f}_\Lambda, \Lambda)_w = \sum_{i=1}^{r} w_i T_i(\widehat{f}_\Lambda, \Lambda)$$

ein Schätzer für $P(\widehat{f}_\Lambda(X_1, \ldots, X_p) = Y)_w$. Die Diskussion im Abschnitt 2.1 zeigt jedoch, dass eine solche Schätzung in der Regel zu optimistische Werte liefert, da die Datenmenge zur Stichprobe Λ in die Festlegung von $\widehat{f}_\Lambda$ eingeflossen ist. Im Folgenden werden zwei Methoden diskutiert mit diesem Problem umzugehen.

Aufteilung der Stichprobe in Trainings- und Test-Menge

Diese Methode wurde beispielhaft bereits im Abschnitt 2.1 vorgestellt: Man unterteilt die vorliegende Stichprobe zufällig in zwei disjunkte Teilmengen

$$\Lambda = \Lambda_{\text{train}} \cup \Lambda_{\text{test}}, \tag{22}$$

ermittelt basierend auf der *Trainingsmenge* Λ_{train} den geschätzen Klassifikator $\widehat{f}_{\Lambda_{\text{train}}}$ und dessen gewichtete Trefferquote $T(\widehat{f}_{\Lambda_{\text{train}}}, \Lambda_{\text{test}})_w$ auf der *Testmenge*. Diese wird dann in der Regel eine weniger optimistische, realistischere Schätzung von $P(\widehat{f}_\Lambda(X_1, \ldots, X_p) = Y)_w$ liefern.

Subsampling

In Verallgemeinerung der obigen Idee könnte man neben Λ weitere Stichproben $\Lambda_1, \ldots, \Lambda_\ell$ aus Ω ziehen und die gewichteten Trefferquoten

$$T(\widehat{f}_\Lambda, \Lambda_j)_w, \ j \in \{1, \ldots, \ell\}$$

betrachten. Deren Mittelwert

$$\overline{T} := \frac{1}{\ell} \sum_{j=1}^{\ell} T(\widehat{f}_\Lambda, \Lambda_j)_w$$

wäre eine realistischere Schätzung der Trefferwahrscheinlichkeit von $\widehat{f}_\Lambda$ und die empirische Varianz

$$\sigma^2 := \frac{1}{\ell - 1} \sum_{j=1}^{\ell} (T(\widehat{f}_\Lambda, \Lambda_j)_w - \overline{T})^2$$

würde die Abhängigkeit der Schätzung der Trefferwahrscheinlichkeit von der gewählten Stichprobe quantifizieren. Die Möglichkeit des Erhebens weiterer Stichproben Λ_j besteht in der Praxis allerdings häufig nicht.

Ist die Stichprobe Λ selbst hinreichend groß, so kann man das Ziehen von Stichproben Λ_j aus Ω simulieren, indem man diese aus Λ selbst zieht: In Abbildung 20 ist das Ergebnis der Anwendung dieses Verfahrens auf die Datenmenge »Mushroom« aus Beispiel 3.2 dargestellt. Es wurde 10 000 mal eine Stichprobe Λ' aus Λ gezogen, jeweils eine Schätzung $\widehat{b}_{n,\Lambda'}$ des naiven Bayes-Klassifikators auf der Basis von Λ' bestimmt und die Trefferquote $T(\widehat{b}_{n,\Lambda'}, \Lambda \setminus \Lambda')$ berechnet. Die Stichproben Λ' enthielten dabei jeweils 50% der Elemente von Λ. Man bezeichnet dieses Vorgehen als *Subsampling*.
Als Ergebnis des Subsampling-Experiments ergeben sich die Werte:

- Trefferquote auf der Gesamtstichprobe: $T(\widehat{b}_{n,\Lambda}, \Lambda) = 0.8481$.
- Mittelwert der Trefferquoten $T(\widehat{b}_{n,\Lambda'}, \Lambda \setminus \Lambda')$: 0.845.
- Standardabweichung der Trefferquoten $T(\widehat{b}_{n,\Lambda'}, \Lambda \setminus \Lambda')$: $0,006$.

Die Differenz zwischen der optimistischen Trefferquote auf der Gesamtstichprobe und dem Mittelwert des Subsampling ist kleiner als eine Standardabweichung.

Abbildung 21 zeigt, dass dies beim 1-NN-Klassifikator f_1 auf der im Abschnitt 2.1 vorgestellten Olivenöl-Datenmenge deutlich anders aussieht. Da die Gesamtstichprobe kleiner ist, wurden nur 1000 Ziehungen vorgenommen.

- Trefferquote auf der Gesamtstichprobe: $T(f_1, \Lambda) = 1.0$ – siehe Feststellung 2.4.
- Mittelwert der Trefferquoten $T(f_1, \Lambda \setminus \Lambda')$: 0.898.
- Standardabweichung der Trefferquoten $T(f_1, \Lambda \setminus \Lambda')$: $0,014$.

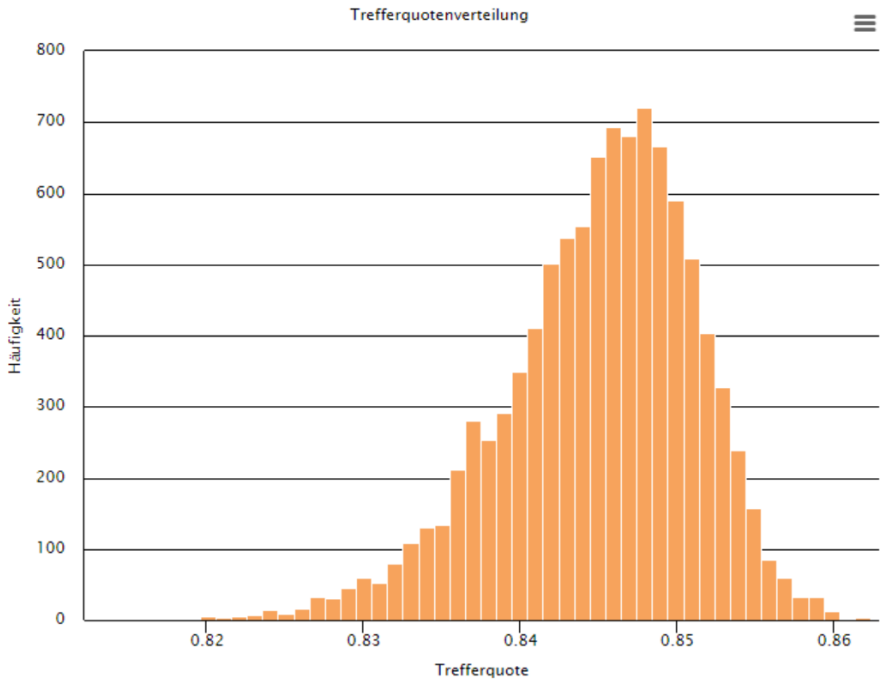


Abbildung 20: Trefferquoten des naiven Bayes-Klassifikators beim Subsampling aus der Datenmenge »Mushroom«

Die Trefferquote $T(f_1, \Lambda) = 1.0$ ist viel zu optimistisch und liegt praktisch »außerhalb« der Trefferquotenverteilung.

Um die Methode des Subsampling möglichst effizient umzusetzen möchte man Schnittmengen zwischen den verschiedenen Stichproben Λ' vermeiden. Dies führt unmittelbar zur Methode der Kreuzvalidierung.

Kreuzvalidierung (Cross Validation)

Das Verfahren der Kreuzvalidierung geht von folgender Situation aus: *Es liegt ein konkreter Algorithmus zur Bestimmung einer Schätzung $\widehat{f}_{\Lambda'} \in \mathbf{F}$ der Lösung eines Klassifikationsproblems (6) auf der Basis einer Stichprobe Λ' vor. Alle im Folgenden auftretenden Klassifikatoren werden mit Hilfe dieses Algorithmus bestimmt.*

Es sei $\Lambda \subseteq \Omega$ eine Stichprobe und $\ell \in \{2, 3, \ldots, |\Lambda|\}$. Ein *$\ell$-Kreuzvalidierungschätzer der gewichteten, totalen Trefferwahrscheinlichkeit von $\widehat{f}_\Lambda$* wird dann nach folgendem Schema berechnet:

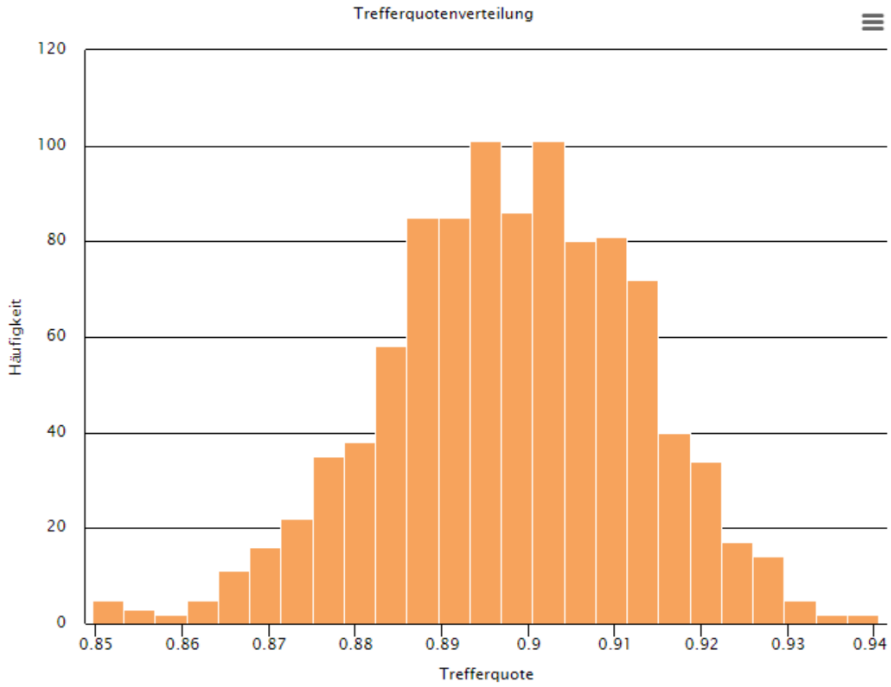


Abbildung 21: Trefferquoten des 1-NN-Klassifikators beim Subsampling aus der Datenmenge »Olive Oils«

1. Unterteile Λ in ungefähr gleichgroße, disjunkte Teilmengen $\Lambda_1, \Lambda_2, \ldots, \Lambda_\ell$.
2. Bestimme die Klassifikatoren $\widehat{f}_{\Lambda\setminus\Lambda_1}, \widehat{f}_{\Lambda\setminus\Lambda_2}, \ldots, \widehat{f}_{\Lambda\setminus\Lambda_\ell}$.
3. Berechne die gewichteten, totalen Trefferquoten $T(\widehat{f}_{\Lambda\setminus\Lambda_k}, \Lambda_k)_w$, $k \in \{1, 2, \ldots, \ell\}$.
4. Dann ist $\widehat{P}_{\ell\mathrm{CV}}(\widehat{f}_\Lambda)_w := \frac{1}{\ell} \sum_{k=1}^{\ell} T(\widehat{f}_{\Lambda\setminus\Lambda_k}, \Lambda_k)_w$ der zu definierende Schätzer.

Die Verwendung von $\ell = 5$ oder $\ell = 10$ ist verbreitet und wird aus Gründen, die hier nicht dargestellt werden, in vielen Fällen auch empfohlen – siehe [FHT, Abschnitt 7.10.1].

Für kleine Stichproben Λ empfiehlt es sich den Wert $\ell = |\Lambda|$ zu nutzen. Die Stichproben Λ_i besitzen dann jeweils nur ein Element. Dieser Spezialfall der Kreuzvalidierung wird auch als *Leaving-One-Out* bezeichnet.

Die Abbildung 22 zeigt links die im Verlauf einer 10-fachen Kreuzvalidierung ermittelten Trefferquoten $T(\widehat{b}_{n,\Lambda\setminus\Lambda_k}, \Lambda_k)$ der Schätzer $\widehat{b}_{n,\Lambda\setminus\Lambda_k}$ des naiven Bayes-Klassifikators auf der Datenmenge »Mushroom«. Der sich ergebende Kreuzvalidierungsschätzer für die totale Trefferwahrscheinlichkeit ist

$$\widehat{P}_{10\mathrm{CV}}(\widehat{b}_{n,\Lambda}) = 0.8481;$$

dies ist genau die Trefferquote von $\widehat{b}_{n,\Lambda}$ auf der Stichprobe Λ.

Man beachte, dass eine Gewichtung der Klassen wie im Beispiel 3.2 nicht vorgenommen wurde, weil diese Funktionalität in Rapidminer nicht zur Verfügung steht.

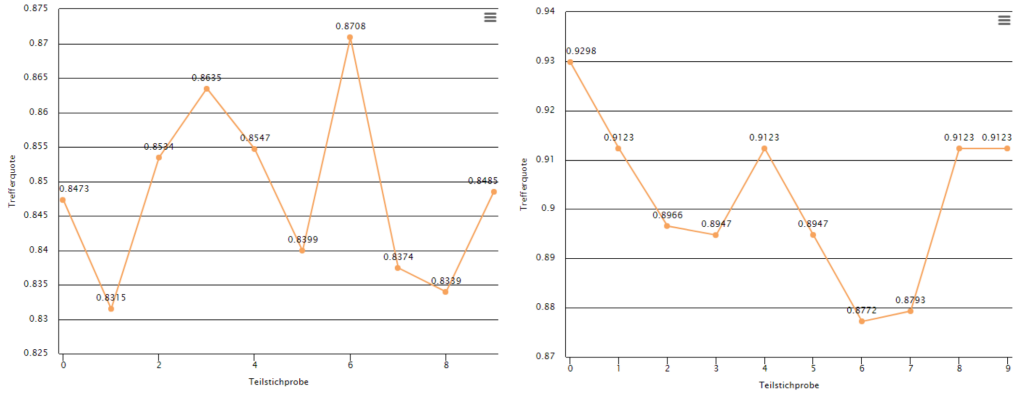


Abbildung 22: Trefferquoten bei 10-fach-Kreuzvalidierung: naiver Bayes-Klassifikator auf der Datenmenge »Mushroom« (links), 1-NN-Klassifikator auf der Datenmenge »Olive Oils« (rechts)

Rechts zeigt Abbildung 22 die Trefferquoten $T(f_{1,\Lambda\setminus\Lambda_k}, \Lambda_k)$ von 1-NN-Klassifikatoren auf der Datenmenge »Olive Oils« im Verlauf einer 10-fachen Kreuzvalidierung. Der sich ergebende Kreuzvalidierungsschätzer für die totale Trefferwahrscheinlichkeit ist

$$\widehat{P}_{10\mathrm{CV}}(f_{1,\Lambda}) = 0.9021.$$

Dem aufmerksamen Leser sollte der folgende Punkt bei der Definition der Kreuzvalidierung aufgefallen sein: Zu jeder Teilmenge $\Lambda_i \subset \Lambda$ wird ein neuer Klassifikator $\widehat{f}_{\Lambda\setminus\Lambda_i}$ ermittelt und dessen Trefferquote auf Λ_i berechnet. Anders als in der ursprünglichen Idee für das Subsampling arbeitet man also nicht mit dem festen Klassifikator $\widehat{f}_\Lambda$, weswegen sich Zweifel daran einstellen können, ob die Kreuzvalidierung wirklich wie behauptet die Trefferquote von $\widehat{f}_\Lambda$ schätzt. Die Ausführungen in [FHT], Abschnitte 7.10 und 7.12 zeigen, dass diese Zweifel berechtigt sind und die Kreuzvalidierung auch bei »großem« ℓ nur die erwartete Trefferquote von $\widehat{f}_\Lambda$ schätzt, wobei auch die Stichprobe Λ als Zufallsgröße aufgefasst wird.

Arten des Subsampling

Zur Durchführung einer Kreuzvalidierung müssen aus der in r disjunkte Klassen unterteilten Stichprobe

$$\Lambda = \Lambda_1 \cup \Lambda_2 \cup \ldots \cup \Lambda_r$$

Teilstichproben $\Lambda' \subset \Lambda$ zufällig gezogen werden. Dies kann man prinzipiell auf zwei verschiedene Weisen tun:

- **Gemischtes Ziehen:** Beim Ziehen von Λ' wird auf die Klassenunterteilung von Λ keine Rücksicht genommen.

  Im Ergebnis können die Klassenanteile $\Lambda' \cap \Lambda_k$ sehr ungleichmäßig ausfallen.

- **Klassenweises Ziehen:** Es werden aus jeder Klasse Λ_k jeweils soviele Elemente zufällig gezogen, dass die entstehende Teilstichprobe Λ' etwa gleiche Klassenanteile wie Λ selbst besitzt.

Bei einer »großen« Stichprobe Λ und hinreichend großen Klassen Λ_k liefert das gemischte Ziehen realistischere Werte als das klassenweise. Ist Λ jedoch klein oder sind einzelne Klassen sehr dünn besetzt, weil die a-priori-Wahrscheinlichkeiten klein sind, so muss man klassenweises Ziehen verwenden um überhaupt brauchbare Kreuzvalidierungsschätzer zu erhalten.

5 Diskriminanzanalyse

Die Bayes-Klassifikatoren in den Sätzen 3.1 und 3.7 besitzen die folgende allgemeine Form:

$$\begin{array}{rcl} f : S_X & \to & \{1, 2, \ldots, r\} \\ x & \mapsto & k \;\Leftrightarrow\; d_k(x) = \max(d_j(x) : j = 1, \ldots, r), \end{array} \tag{23}$$

wobei $d_j : S_X \to \mathbb{R}$ gewisse mit dem betrachteten Klassifikationsproblem zusammenhängende Funktionen sind, nämlich gewichtete a-posteriori-Wahrscheinlichkeiten oder gewichtete Wahrscheinlichkeitsdichten. Es liegt nahe diese Tatsache zu nutzen, um alternative Lösungsansätze für das Klassifikationsproblem zu finden. Die resultierenden Verfahren laufen in der Literatur unter dem Namen *diskriminanzanalytische Verfahren.*

5.1 Grundlegendes

In diesem und den folgenden Abschnitten wird die folgende Situation betrachtet: Die Merkmale $X_1, \ldots, X_p$ sind alle reellwertig, also $S_k \subseteq \mathbb{R}$, womit $S_X \subseteq \mathbb{R}^p$ gilt. Weiter sei $\mathbf{F}$ eine Klasse von Funktionen $d : D \to \mathbb{R}$, wobei $S_X \subseteq D$ gelte. Gesucht sind nun r Funktionen

$$d_1, d_2, \ldots, d_r \in \mathbf{F}$$

derart, dass die erwartete Trefferwahrscheinlichkeit des durch diese Funktionen festgelegten Klassifikators (23) möglichst hoch ist. Die Funktionen d_j bezeichnet man als *Diskriminanzfunktionen (für das vorliegende Klassifikationsproblem).*

Der Zwei-Klassen-Fall

Im Fall zweier Klassen ($r = 2$) sind zwei Diskriminanzfunktionen $d_1, d_2 \in \mathbf{F}$ zu bestimmen. Der zugehörige Klassifikator (23) kann dann mit Hilfe der Funktion

$$d : D \to \mathbb{R},\; x \mapsto d_1(x) - d_2(x) \tag{24}$$

dargestellt werden:

$$f(x) = 1 \;\Leftrightarrow\; d(x) > 0. \tag{25}$$

Eine noch kompaktere Darstellung ergibt sich, wenn man statt der Klassenlabels 1 und 2 die Labels 1 und -1 verwendet:

$$f(x) = \operatorname{sgn}(d(x)), \tag{26}$$

wobei sgn die Vorzeichenfunktion (Signum-Funktion) bezeichnet.

Die Funktion d wird in der Literatur inkonsequenterweise ebenfalls als Diskriminanzfunktion bezeichnet. Man beachte dabei, dass d nicht notwendigerweise im Modellraum $\mathbf{F}$ liegt; dies gilt aber beispielsweise dann, wenn $\mathbf{F}$, wie häufig der Fall, ein reeller Vektorraum ist.

Die Nullstellenmenge

$$H := d^{-1}(0) = \{x \in D : d(x) = 0\}$$

»trennt« die beiden Klassen $f^{-1}(-1)$ und $f^{-1}(1)$ im Raum $\mathbb{R}^p$. Ist D eine offene Teilmenge des $\mathbb{R}^p$ und besitzt d hinreichend »gute« Eigenschaften wie zum Beispiel stetige Differenzierbarkeit mit $f'(x) \neq 0$ für alle $x \in d^{-1}(0)$, so ist nach dem Satz über implizite Funktionen H ein Stück einer Hyperfläche im $\mathbb{R}^p$. Diese Voraussetzungen sind im einfachsten Fall für die Klasse $\mathbf{F} = \mathrm{Aff}\,(\mathbb{R}^p, \mathbb{R})$ der affinen Funktionen $d : \mathbb{R}^p \to \mathbb{R}$ erfüllt. Jede affine Funktion lässt sich in der Form

$$d(x) = \langle a, x\rangle + b, \quad a \in \mathbb{R}^p,\ b \in \mathbb{R}, \tag{27}$$

schreiben, wobei $\langle \cdot, \cdot \rangle$ das Standardskalarprodukt des $\mathbb{R}^p$ ist. Ist $a \neq 0$, so ist H eine Hyperebene mit dem Normalenvektor a.

Zusammengesetzte Klassifikatoren

Prinzipiell kann der Fall von $r > 2$ Klassen durch wiederholte Zwei-Klassen-Klassifikation behandelt werden. Genauer:

1. Man zerlege die vorliegende Stichprobe Λ in zwei disjunkte, annähernd gleich große Teilmengen, die selbst Vereinigungen von Klassen sind:

$$\Lambda = \Delta_1 \cup \Delta_2,\ \Delta_1 = \bigcup_{i=1}^{s} \Lambda_{k_i},\ \Delta_2 = \bigcup_{i=s+1}^{r} \Lambda_{k_i}.$$

2. Man bestimme einen Klassifikator für das durch $\Lambda = \Delta_1 \cup \Delta_2$ gegebene Zweiklassenproblem.

3. Man wiederhole gegebenenfalls die Schritte 1 und 2 mit den Stichprobenteilmengen Δ_1 und Δ_2 solange, bis die entstehenden Teilmengen selbst Klassen Λ_k sind.

4. Man setze die ermittelten Klassifikatoren zu einem Gesamtklassifikator zusammen.

Im Fall $r = 4$ könnte das obenstehende Verfahren zum Beispiel wie folgt ablaufen: In der Klassenzerlegung

$$\Lambda = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3 \cup \Lambda_4$$

sind die Teilmengen $\Delta_1 := \Lambda_1$ und $\Delta_2 := \Lambda_2 \cup \Lambda_3 \cup \Lambda_4$ etwa gleich groß. Man bestimmt daher eine Diskriminanzfunktion d_1 derart, dass

$$f_1(x) = 1 \Leftrightarrow d_1(x) > 0$$

gilt.

Weiter sind die Teilmengen $\Gamma_1 := \Lambda_2 \cup \Lambda_3$ und $\Gamma_2 := \Lambda_4$ ungefähr gleich groß. Zu dem Klassifikationsproblem $\Gamma = \Gamma_1 \cup \Gamma_2$ bestimmt man einen Klassifikator der Form

$$f_2(x) = 1 \Leftrightarrow d_2(x) > 0$$

mit einer Diskriminanzfunktion d_2.

Schließlich bestimmt man zu dem Klassifikationsproblem $\Theta := \Theta_1 \cup \Theta_2$, $\Theta_1 := \Lambda_2$, $\Theta_2 := \Lambda_3$ eine Diskriminanzfunktion d_3 derart, dass

$$f_3(x) = 1 \Leftrightarrow d_3(x) > 0$$

gilt.

Der zugehörige zusammengesetzte Klassifikator für das Vier-Klassen-Problem ist dann:

$$f(x) = \begin{cases} 1 & \text{falls } d_1(x) > 0 \\ 2 & \text{falls } d_2(x) > 0 \wedge d_3(x) > 0 \\ 3 & \text{falls } d_2(x) > 0 \wedge d_3(x) < 0 \\ 4 & \text{falls } d_2(x) < 0. \end{cases}$$

5.2 Lineare Diskriminanzanalyse nach Fisher

Ein Verfahren zur Schätzung affiner Diskriminanzfunktionen anhand der Daten einer Stichprobe wurde im Jahr 1936 von dem Statistiker Ronald Aylmer Fisher vorgeschlagen – siehe [Fis]. Dieses Verfahren kann geometrisch motiviert werden, besitzt aber auch statistische Optimalitätseigenschaften.

Das Verfahren im Fall $r = 2$ aus geometrischer Sicht
Entsprechend der in diesem Abschnitt betrachteten Situation seien

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$$

Abbildung 23: Ronald Aymler Fisher: britischer Statistiker, Genetiker, Evolutionstheoretiker, 1890 – 1962; links: im Jahr 1913, rechts: Portrait (Leontine Tintner Camprubi)

die Daten zu einer Stichprobe Λ aus einer in zwei Klassen zerfallenden Population $\Omega = \Omega_1 \cup \Omega_2$.

Gesucht ist eine Diskriminanzfunktion d im Sinn der Formel (25); diese soll affin sein:

$$d(x) = \langle x, a \rangle + b.$$

In dem hier vorgestellten Verfahren bestimmt man zunächst den Vektor $a \in \mathbb{R}^p$ und dann die reelle Konstante b.

Die Klassifikationsregel (25) ändert sich nicht, wenn man d durch die ebenfalls affine Funktion λd, $\lambda \in \mathbb{R}^{>0}$, ersetzt. Man kann also ohne Einschränkung der Allgemeinheit $\|a\| = 1$ annehmen.

Interpretiert man die Klassifikationsregel (25) geometrisch, so trennt die Hyperebene $H = d^{-1}(0)$ die beiden Punktfamilien (mehrfach vorkommende Punkte $x^{(i)}$ werden mehrfach gezählt!)

$$C_1 := (x^{(i)} : y^{(i)} = 1), \quad C_2 := (x^{(i)} : y^{(i)} = 2)$$

möglichst »optimal«. Fisher betrachtet die Hyperebene H als »optimal«, falls die auf H senkrecht stehende Gerade $g = \mathbb{R}a$ die folgenden beiden Eigenschaften besitzt: Ist

$$\pi_g : \mathbb{R}^p \to g$$

die orthogonale Projektion auf g, so sollen

1. Die Mittelwerte der Familien $\pi_g(C_1)$ und $\pi_g(C_2)$ möglichst großen Abstand voneinander haben.

2. Die Varianzen der Familien $\pi_g(C_1)$ und $\pi_g(C_2)$ möglichst klein sein.

Die orthogonale Projektion π_g kann konkret durch ein Skalarprodukt ausgedrückt werden:

$$\pi_g(x) = \langle x, a\rangle a,$$

eine Tatsache, die zur Bestimmung von a genutzt wird.

In der Abbildung 24 sind für mit einem Zufallsgenerator erzeugte, normalverteilte Daten $C_1, C_2 \subset \mathbb{R}^2$ bestehend aus jeweils 500 Samples die Verteilungen der Werte $\langle x, a\rangle$ für drei verschiedene Vektoren a dargestellt.

Für den Mittelwert der Familie C_k gilt

$$\overline{x}_k = \frac{1}{|C_k|} \sum_{y^{(i)}=k} x^{(i)}$$

und für das Quadrat des euklidischen Abstands ihrer Projektionen

$$\begin{aligned} \|\pi_g(\overline{x}_1) - \pi_g(\overline{x}_2)\|^2 &= \|\langle \overline{x}_1, a\rangle a - \langle \overline{x}_2, a\rangle a\|^2 \\ &= \langle \overline{x}_1 - \overline{x}_2, a\rangle^2 \\ &= a^T(\overline{x}_1 - \overline{x}_2)(\overline{x}_1 - \overline{x}_2)^T a \\ &= a^t B a \end{aligned}$$

mit der symmetrischen Matrix

$$B := (\overline{x}_1 - \overline{x}_2)(\overline{x}_1 - \overline{x}_2)^T \in \mathbb{R}^{p\times p}.$$

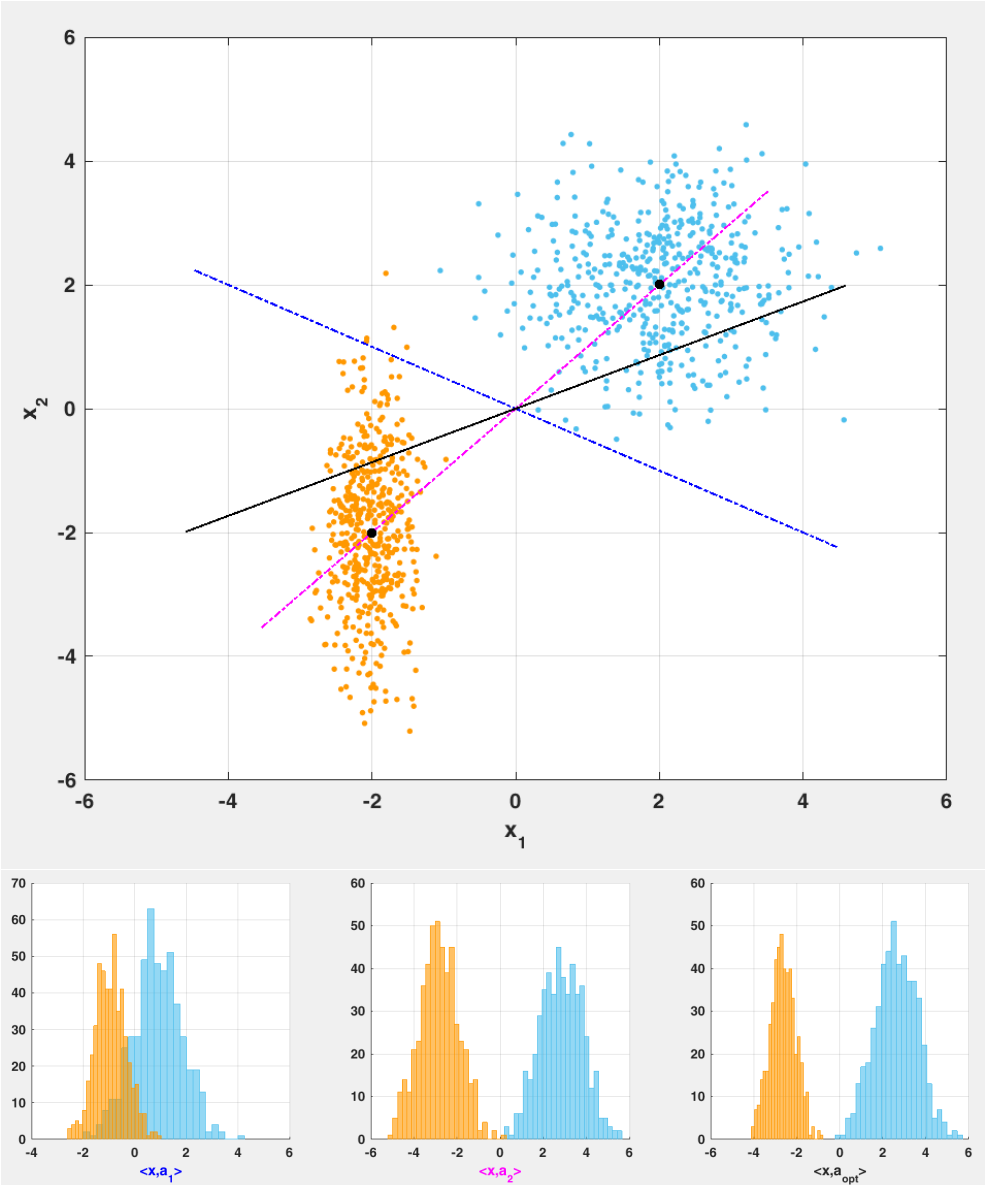


Abbildung 24: Orthogonale Projektion von Daten im Zwei-Klassen-Fall

Um die Varianz der projizierten Familie $\pi_g(C_k)$ zu berechnen, beachte man zunächst, dass wegen der Linearität der Projektion π_g der Mittelwert der projizierten Familie $\pi_g(C_k)$ gleich dem projizierten Mittelwert $\pi_g(\overline{x}_k)$ ist. Es folgt:

$$\begin{aligned}
\sigma_k^2 &= \tfrac{1}{|C_k|} \sum_{y^{(i)}=k} \|\langle x^{(i)}, a\rangle a - \langle x^{(i)}, a\rangle a\|^2 \\
&= \tfrac{1}{|C_k|} \sum_{y^{(i)}=k} (\langle x^{(i)}, a\rangle - \langle x^{(i)}, a\rangle)^2 \\
&= \tfrac{1}{|C_k|} \sum_{y^{(i)}=k} \langle x^{(i)} - \overline{x}_k, a\rangle^2 \\
&= \tfrac{1}{|C_k|} \sum_{y^{(i)}=k} a^t (x^{(i)} - \overline{x}_k)(x^{(i)} - \overline{x}_k)^t a \\
&= a^t (\tfrac{1}{|C_k|} \sum_{y^{(i)}=k} (x^{(i)} - \overline{x}_k)(x^{(i)} - \overline{x}_k)^t) a \\
&= a^t W_k a,
\end{aligned}$$

mit der symmetrischen Matrix

$$W_k := \frac{1}{|C_k|} \sum_{y^{(i)}=k} (x^{(i)} - \overline{x}_k)(x^{(i)} - \overline{x}_k)^t.$$

Die Gerade g und damit der Vektor a müssen zwei Forderungen erfüllen, die möglicherweise nicht simultan erfüllbar sind, sodass eine kombinierte und ausbalancierte Betrachtung der Forderungen nötig ist. Dies kann man etwa mit Hilfe der Funktion

$$R(a) := \frac{a^t B a}{a^t W_1 a + a^t W_2 a} = \frac{a^t B a}{a^t W a},\ W := W_1 + W_2 \tag{28}$$

erreichen, indem man deren Maxima betrachtet: Der Grund hierfür wird durch die einfache Beobachtung geliefert, dass die Funktionswerte von $R(a)$ umso größer werden, je weiter die projizierten Mittelwerte auseinanderliegen und je kleiner die Varianzsumme der projizierten Daten ist.

Liegen die Stichprobendaten nicht in einem echten Untervektorraum $U \subset \mathbb{R}^p$, so ist die Matrix W invertierbar und R ist daher auf der offenen Menge $\mathbb{R}^p \setminus 0$ definiert. Dies ist der Normalfall, der im Weiteren vorausgesetzt wird.

R besitzt ein Maximum: Es gilt $R(\lambda a) = R(a)$ für alle $\lambda \neq 0$, womit ein Maximum von R gegebenenfalls bereits in der Einheitssphäre $S_p := \{a \in \mathbb{R}^p : \|a\| = 1\}$ angenommen wird. Diese ist kompakt und R ist als gebrochenrationale Funktion stetig, folglich besitzt R ein Maximum.

Die Funktion R ist als gebrochenrationale Funktion sogar differenzierbar, womit man das Maximum mittels Differentialrechnung bestimmen kann.

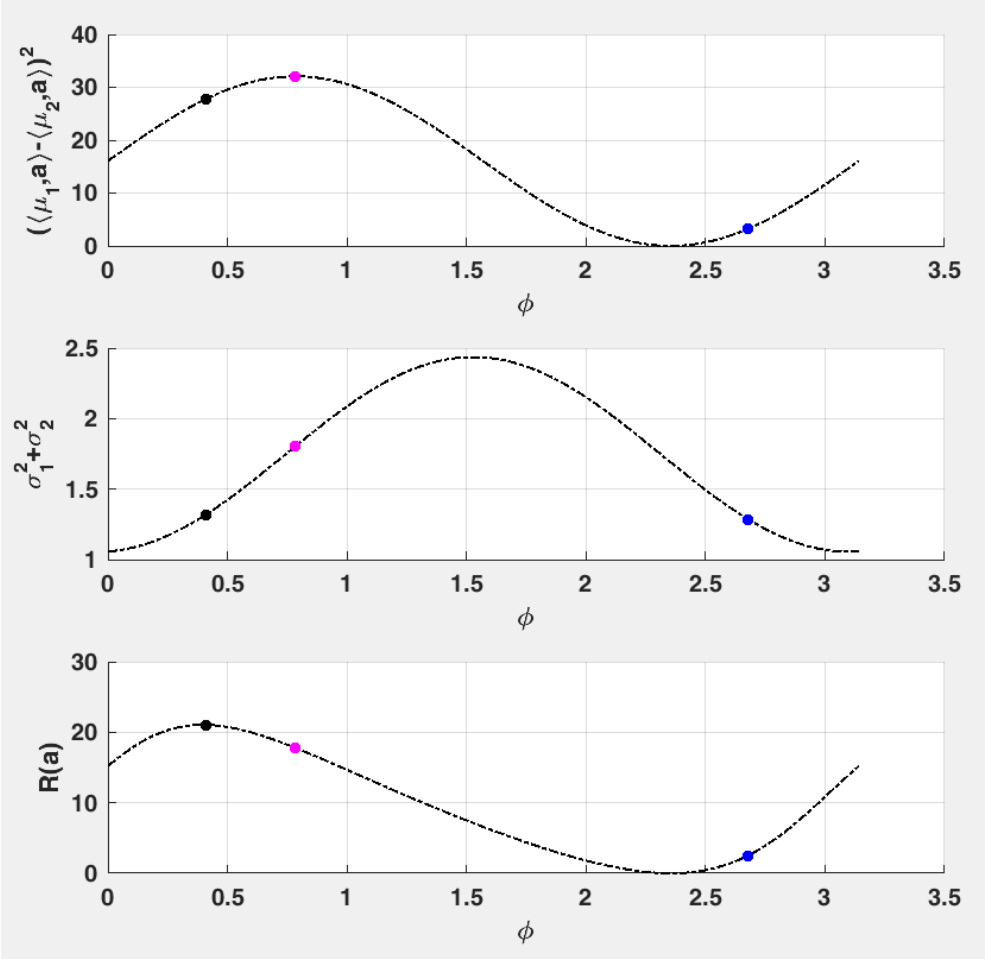


Abbildung 25: Die Fisher-Funktion (28) für die Daten aus Abbildung 24
Anmerkung: Statt des Vektors $a \in \mathbb{R}^2$ mit $\|a\| = 1$ ist der Winkel $\phi \in [0, \pi]$ mit $a = (\cos(\phi), \sin(\phi))^t$ auf der Rechtsachse abgetragen.

Die Abbildung 25 zeigt Werte zur Funktion $R(a)$ für die in der Abbildung 24 gezeigten Daten in Abhängigkeit vom Winkel $\phi \in [0, \pi]$, den die Gerade g mit der x_1-Achse bildet. Dabei zeigt das obere der drei Diagramme den Zähler a^tBa und das mittlere den Nenner a^tWa von $R(a)$; das untere Diagramm schließlich stellt die Funktion R selbst dar. Der schwarz markierte Punkt auf dem Graphen von R markiert das Maximum der Funktion, an dessen Position durchaus nicht der maximale Abstand der projizierten Mittelwerte angenommen wird oder der minimale Wert der Summe der Varianzen der projizierten Daten.

Eine notwendige Bedingung für das Vorliegen eines Maximums von R bei $a^* \in \mathbb{R}^p$ ist

$$R'(a^*) = 0;$$

es sind also die partiellen Ableitungen $\frac{\partial R}{\partial a_k}$, $a = (a_1, \ldots, a_p)$, zu berechnen.

Ist $C = (c_{ij}) \in \mathbb{R}^{p \times p}$ eine symmetrische Matrix, so gilt

$$\begin{aligned} \frac{\partial}{\partial a_k}(a^tCa) &= \frac{\partial}{\partial a_k}(\sum_{i=1}^{p}\sum_{j=1}^{p} c_{ij}a_ia_j) \\ &= \sum_{i \neq k} c_{ik}a_ia_k) + \sum_{j \neq k} c_{kj}a_ka_j + 2c_{kk}a_k \\ &= 2a^tc_k, \end{aligned}$$

wobei c_k die k-te Spalte der Matrix C ist, und man die Symmetrie von C benutzt. Es folgt

$$\begin{aligned} \frac{\partial R}{\partial a_k} &= \frac{\partial}{\partial a_k}\left(\frac{a^tBa}{a^tWa}\right) \\ &= \frac{2a^tb_k(a^tWa) - (a^tBa)2a^tw_k}{(a^tWa)^2} \end{aligned}$$

also

$$R'(a) = \frac{2(a^tWa)a^tB - 2(a^tBa)a^tW}{(a^tWa)^2}.$$

Die Extremwertbedingung läuft also auf die Gleichung

$$(a^tWa)a^tB - (a^tBa)a^tW = 0$$

hinaus. Wegen $a^tB = a^t(\overline{x}_1 - \overline{x}_2)(\overline{x}_1 - \overline{x}_2)^t$ kann man diese Gleichung durch $a^t(\overline{x}_1 - \overline{x}_2)$ teilen, sofern diese Zahl nicht 0 ist. Das wäre genau dann der Fall, wenn a orthogonal zu $\overline{x}_1 - \overline{x}_2$ ist. In diesem Fall werden die Punkte $\overline{x}_1$ und $\overline{x}_2$ auf denselben Punkt projiziert, was der Bedingung 1 widerspricht. In für die Anwendung relevanten Situationen wird dieser Fall also nicht vorliegen. Es folgt

$$(a^tWa)(\overline{x}_1 - \overline{x}_2)^t - (\overline{x}_1 - \overline{x}_2)^taa^tW = 0$$

also

$$Wa = \frac{(a^t W a)}{a^t(\overline{x}_1 - \overline{x}_2)}(\overline{x}_1 - \overline{x}_2).$$

Da die Matrix W nach Voraussetzung invertierbar ist, kann man die letzte Gleichung zu

$$a = \frac{(a^t W a)}{a^t(\overline{x}_1 - \overline{x}_2)} W^{-1}(\overline{x}_1 - \overline{x}_2)$$

umformen. Man beachte, dass die so erhaltene Gleichung *keine* explizite Formel für a ist, da der reelle Faktor $\frac{(a^t W a)}{a^t(\overline{x}_1 - \overline{x}_2)}$ selbst von a abhängt und unbekannt ist. Allerdings muss a auch nur bis auf reelle Vielfache $\lambda \neq 0$ bestimmt werden, da man dann durch Normieren das gesuchte a erhält. Aus der letzten Gleichung folgt also:

$$a^* = \frac{W^{-1}(\overline{x}_1 - \overline{x}_2)}{\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|}. \tag{29}$$

Es bleibt den konstanten Koeffizienten b der gesuchten affinen Diskriminanzfunktion zu berechnen: Geometrisch ist es naheliegend die Klassengrenze H durch den Mittelwert $\frac{1}{2}(\langle a^*, \overline{x}_1\rangle + \langle a^*, \overline{x}_2\rangle)a^*$ der projizierten Mittelwerte der Klassen C_1 und C_2 laufen zu lassen. In diesem Fall gilt:

$$d(\frac{1}{2}(\langle a^*, \overline{x}_1\rangle + \langle a^*, \overline{x}_2\rangle)a^*) = \langle a^*, \frac{1}{2}(\langle a^*, \overline{x}_1\rangle + \langle a^*, \overline{x}_2\rangle)a^*\rangle + b = 0,$$

woraus sich die Gleichung

$$b = -\langle a^*, \frac{1}{2}(\overline{x}_1 + \overline{x}_2)\rangle \tag{30}$$

ergibt. Man beachte jedoch, dass es im Fall stark unterschiedlicher Klassengrößen sinnvoll sein kann den Parameter b so zu wählen, dass die erwartete Trefferquote möglichst hoch ist.

Um die Klassifikationsregel angeben zu können, die sich aus Fishers Diskriminanzfunktion

$$d_F : \mathbb{R}^p \to \mathbb{R},\ x \mapsto \langle a^*, x\rangle + b \tag{31}$$

ergibt, ist noch das Vorzeichen der Werte von $d_F(x)$ für die beiden Klassen zu ermitteln: Gemäß der geometrischen Idee, die hinter der Konstruktion von d_F steht, muss der Mittelwert $\overline{x}_k$ der Familie C_k durch die zu d_F gehörende Klassifikationsregel jeweils der Klasse k zugeordnet werden. Es gilt

$$\begin{aligned} d_F(\overline{x}_k) &= \tfrac{1}{\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|}\langle W^{-1}(\overline{x}_1 - \overline{x}_2), \overline{x}_k\rangle - \langle W^{-1}(\overline{x}_1 - \overline{x}_2), \tfrac{1}{2}(\overline{x}_1 + \overline{x}_2)\rangle \\ &= \tfrac{1}{\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|}\langle W^{-1}(\overline{x}_1 - \overline{x}_2), \tfrac{1}{2}(\overline{x}_k - \overline{x}_j)\rangle \\ &= \tfrac{1}{\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|}\tfrac{1}{2}(\overline{x}_k - \overline{x}_j)^t W^{-1}(\overline{x}_1 - \overline{x}_2), \end{aligned}$$

wobei $j \neq k$ ist. Es folgt $d_F(\overline{x}_1) > 0$ und $d_F(\overline{x}_2) < 0$ und damit Fishers Klassifikationsregel

$$f_F(x) = 1 \Leftrightarrow d_F(x) > 0 \tag{32}$$

bei Verwendung der Klassenlabels $\{1, 2\}$.

Eine bestechende Eigenschaft von Fishers Klassifikationsregel ist die Tatsache, dass man die Werte $d_F(x)$ der Diskriminanzfunktion geometrisch deuten kann:

FESTSTELLUNG 5.1: *Für Fishers Diskriminanzfunktion d_F gilt: Der Wert $|d_F(x)|$ ist der euklidische Abstand des Punktes $x \in \mathbb{R}^p$ von der klassentrennenden Hyperebene $H = d_F^{-1}(0)$.*

BEWEIS: Der Abstand von x von H ist per Definition gleich $\|x - x_0\|$, wobei $x_0 \in H$ die orthogonale Projektion von x auf H ist. Da die Gerade $g = \mathbb{R}a_0$ senkrecht auf H steht, gilt

$$\|x - x_0\| = \|\pi_g(x) - x_s\|,$$

wobei $x_s = H \cap g$ gilt. Nach Konstruktion der Ebene H ist

$$x_s = \frac{1}{2}(\overline{x}_1 + \overline{x}_2) = \frac{1}{2}(\langle a^*, \overline{x}_1\rangle + \langle a^*, \overline{x}_2\rangle)a^*$$

und daher

$$\begin{aligned} \|\pi_g(x) - x_s\| &= \|\langle a^*, x\rangle a^* - \tfrac{1}{2}(\langle a^*, \overline{x}_1\rangle + \langle a^*, \overline{x}_2\rangle)a^*\| \\ &= |d_F(x)|. \end{aligned}$$

□

Der Abstand eines Punktes $x \in S_X$ von der klassentrennenden Hyperebene H kann heuristisch als »Sicherheit der Klassifikationsentscheidung« gedeutet werden, was der obigen Feststellung eine praktische Bedeutung gibt.

BEISPIEL 5.2 (Identifikation von Glas): Wir betrachten die Datenmenge »Glass Identification« aus dem UCI-Machine-Learning-Repository [UCI]. Sie umfasst die chemische Zusammensetzung von 214 Glassamples beschrieben durch die Gewichtsanteile der Oxide von sechs chemischen Elementen, sowie dem Brechungsindex des jeweiligen Glases als siebtem Merkmal. Die Samples sind in sieben Klassen eingeteilt, von denen im vorliegenden Beispiel nur zwei betrachtet werden:

- Klasse 1: Gebäudeglas »float processed« (70 Samples),
- Klasse 2: Gebäudeglas »non float processed« (76 Samples).

Von den Merkmalen zur chemischen Zusammensetzung werden drei genutzt:

- X_1: Gewichtsanteil von Natriumoxid (Na_2O),
- X_2: Gewichtsanteil von Aluminiumoxid (Al_2O_3),
- X_3: Gewichtsanteil von Siliziumoxid (SiO_2).

Die Wahl der Merkmale wird unter anderem durch die folgenden Hintergrundinformationen begründet: Die Datenmenge wurde ursprünglich zum Testen einer kommerziellen Software namens *Beagle* erhoben, mit deren Hilfe die Herkunft von Glasscherben (Glasart) im Rahmen von kriminologischen Ermittlungen bestimmt werden konnte – die Software scheint aktuell nicht mehr verfügbar zu sein. Je nach Art besteht Glas chemisch betrachtet neben dem Hauptbestandteil Siliziumdioxid aus verschiedenen Metalloxiden und Metallen, welche die optischen und mechanischen Eigenschaften des Glases bestimmen. Glas für die Verwendung in Gebäuden wird häufig mit einem als »float processing« bezeichneten Verfahren hergestellt, bei dem das geschmolzene Glas zeitweilig auf einem Bett aus geschmolzenem Zinn schwimmt, wodurch Platten unterschiedlicher und sehr gleichmäßiger Dicke hergestellt werden können. Die Eignung einer Glasart für das »float processing« hängt von dessen chemischer Zusammensetzung ab, die oben genannten drei Oxide spielen dabei eine herausgehobene Rolle.

Vor der Analyse werden die Ausprägungen

$$x^{(j)} = (x_1^{(j)}, x_2^{(j)}, x_3^{(j)})$$

aller drei Merkmale standardisiert, also mittels

$$\tilde{x}^{(j)} := (\frac{x_1^{(j)} - \overline{x}_1}{\sigma_1}, \frac{x_2^{(j)} - \overline{x}_2}{\sigma_2}, \frac{x_3^{(j)} - \overline{x}_3}{\sigma_3})$$

transformatiert, wobei $\overline{x}_i$ der Mittelwert und σ_i die Standardabweichung der Ausprägungen des Merkmals X_i in der Stichprobe Λ sind.

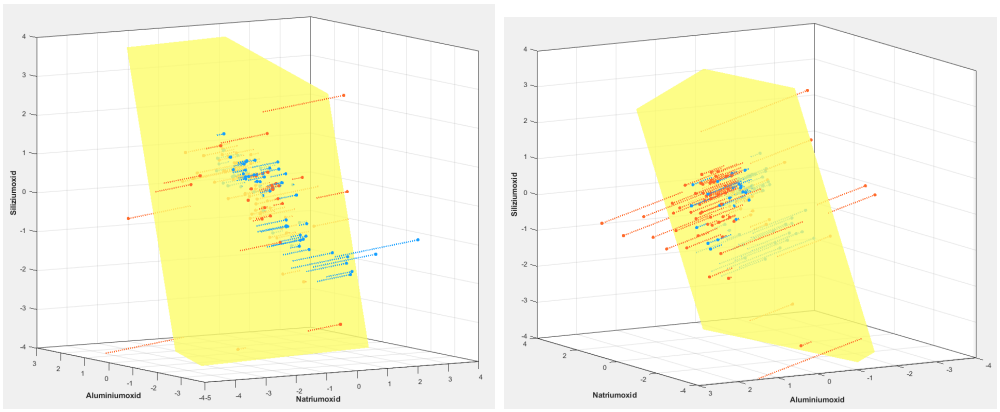


Abbildung 26: Diskriminanzanalyse der Datenmenge »Glass Identification«

In Abbildung 26 sind die transformierte Datenmenge und die nach dem Verfahren von Fisher ermittelte, klassentrennende Ebene in zwei Ansichten dargestellt. Die Samples der Klasse 1 erscheinen hierbei blau, die der Klasse 2 orangefarben. Die punktierten Linien deuten jeweils den Abstand zur trennenden Ebene H an, die durch die Gleichung

$$0.127X_1 - 0.958X_2 + 0.256X_3 - 0.0157 = 0 \tag{33}$$

gegeben ist. Die Abbildung 27 zeigt die klassenweise Verteilung der Abstände von H in einem Histogramm.

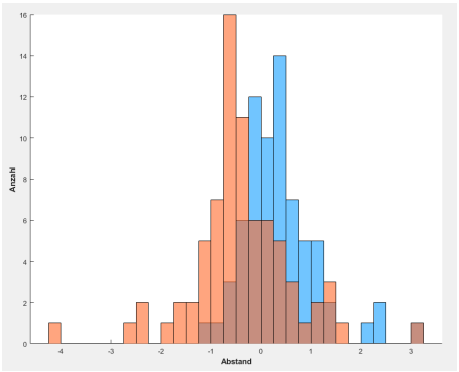


Abbildung 27: Diskriminanzanalyse der Datenmenge »Glass Identification«

Der erhebliche Überlapp der beiden Klassen im Histogramm deutet bereits darauf hin, dass die Fisher-Methode im vorliegenden Fall keine befriedigende Klassentrennung liefert, was durch die Schätzung der zu erwartenden Trefferquote für den zur Gleichung (33) gehörenden Klassifikator f_F (32) bestätigt wird:

$$T_{\mathrm{LOO}}(f_F, \Lambda) = 0.65.$$

Wegen der vergleichsweise niedrigen Samplezahlen wurde die Schätzung mit dem Leaving-One-Out-Verfahren durchgeführt. ◇

Beispiel 5.3 (Forts. von Beispiel 3.8): Mit diesem Beispiel wird einerseits die Anwendung von Fishers Methode demonstriert und andererseits die Erstellung eines zusammengesetzten Klassifikators wie im Abschnitt 5.1 beschrieben.

Die drei Weizenarten sind in der Datenmenge »Seeds« gleich stark vertreten. Betrachtet man die Verteilung der Daten (Abbildung 28), dieses Mal allerdings in Bezug auf die Merkmale

$$\begin{array}{rcl} X_1 & := & \text{A (Querschnittsfläche des Korns)}, \\ X_2 & := & \text{LKG (Länge der Korneinkerbung)}, \end{array}$$

so erscheint es sinnvoll zunächst eine Diskriminanzfunktion $d_{2,\{1,3\}}$ zur Trennung der Klasse $\Delta_1 := \Lambda_2$ (Weizensorte Rosa) vom Rest $\Delta_2 := \Lambda_1 \cup \Lambda_3$ zu ermitteln und danach die Klassen Λ_1 (Kama) und Λ_3 (Canadian) durch eine Diskriminanzfunktion $d_{1,3}$ zu trennen.

Die Formeln (29) und (30) liefern die Diskriminanzfunktionen

$$\begin{array}{rcl} d_{2,\{1,3\}}(x) & = & -0.106x_1 - 0.994x_2 + 7.201 \\ d_{1,3}(x) & = & -0.328x_1 + 0.945x_2 - 0.526 \end{array}$$

– siehe die Abbildung 28.

Der resultierende zusammengesetzte Klassifikator ist

$$f(x) := \begin{cases} 1 & \text{falls } d_{2,\{1,3\}}(x) < 0 \wedge d_{1,3}(x) > 0 \\ 2 & \text{falls } d_{2,\{1,3\}}(x) > 0 \\ 3 & \text{falls } d_{2,\{1,3\}}(x) < 0 \wedge d_{1,3}(x) < 0. \end{cases}$$

◇

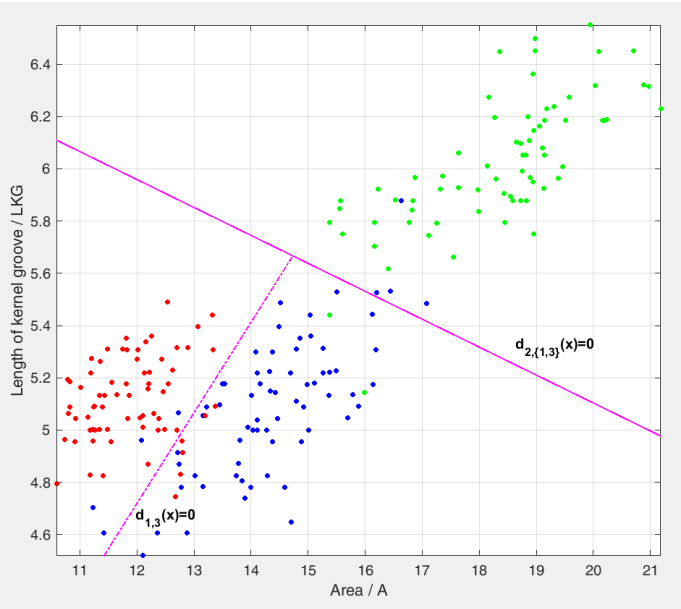


Abbildung 28: Fishers Diskriminanzfunktionen für die Datenmenge »Seeds« (Beispiel 5.3), Klassen: 1: blau, 2: grün, 3: rot.

Zur Optimalität von Fishers Diskriminanzfunktion

Die aus linearen Diskriminanzfunktionen nach Fisher gewonnenen Klassifikatoren können in bestimmten Fällen als Schätzungen von Bayes-Klassifikatoren gedeutet werden und besitzen daher entsprechende stochastische Optimalitätseigenschaften.

Wir betrachten das allgemeine Klassifikationsproblem in der in Satz 3.7 vorausgesetzten Form und nehmen außerdem den Zwei-Klassen-Fall an. Sind dann die Zufallsvariablen $Z_1 := (X_1, \ldots, X_p)|_{\Omega_1}$ und $Z_2 := (X_1, \ldots, X_p)|_{\Omega_2}$ normalverteilt und besitzen dieselbe Kovarianzmatrix

$$\Sigma = \Sigma_1 = \Sigma_2, \tag{34}$$

so vereinfacht sich die definierende Gleichung für den zugehörigen Bayes-Klassifikator (20) und (21) wie folgt:

$$\begin{aligned} T(x) &= (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) + \ln(\tfrac{\det(\Sigma_1)}{\det(\Sigma_2)}) - 2\ln(\tfrac{w_1}{w_2}) \\ &= (x-\mu_1)^t\Sigma^{-1}(x-\mu_1) - (x-\mu_2)^t\Sigma^{-1}(x-\mu_2) - 2\ln(\tfrac{w_1}{w_2}) \\ &= x^t\Sigma^{-1}x - 2\mu_1^t\Sigma^{-1}x + \mu_1^t\Sigma^{-1}\mu_1 - x^t\Sigma^{-1}x + 2\mu_2^t\Sigma^{-1}x - \mu_2^t\Sigma^{-1}\mu_2 - 2\ln(\tfrac{w_1}{w_2}) \\ &= 2(\mu_2-\mu_1)^t\Sigma^{-1}x + \mu_1^t\Sigma^{-1}\mu_1 - \mu_2^t\Sigma^{-1}\mu_2 - 2\ln(\tfrac{w_1}{w_2}). \end{aligned}$$

Nun sind die Matrizen W_1 und W_2 nach Definition und wegen der Voraussetzung (34) beide Schätzer für die Kovarianzmatrix Σ, daher ist W ein Schätzer für 2Σ:

$$\widehat{\Sigma} = \frac{1}{2}W.$$

Ersetzt man in der Gleichung für T alle Größen durch ihre Schätzer, so ergibt sich die Schätzung

$$\begin{aligned}
\widehat{d}(x) &= 4(\overline{x}_2 - \overline{x}_1)^t W^{-1} x + 2\overline{x}_1^t W^{-1}\overline{x}_1 - 2\overline{x}_2^t W^{-1}\overline{x}_2 - 2\ln(\tfrac{w_1}{w_2}) \\
&= 2(\langle 2W^{-1}(\overline{x}_2 - \overline{x}_1), x\rangle + \overline{x}_1^t W^{-1}\overline{x}_1 - \overline{x}_2^t W^{-1}\overline{x}_2 - \ln(\tfrac{w_1}{w_2})) \\
&= 2(\langle -2\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|a^*, x\rangle + \overline{x}_1^t W^{-1}\overline{x}_1 - \overline{x}_2^t W^{-1}\overline{x}_2 - \ln(\tfrac{w_1}{w_2}))
\end{aligned}$$

für d, wobei man die Symmetrie der Matrix W und damit ihrer Inversen W^{-1} benutzt. Es gilt weiter

$$\begin{aligned}
-2\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|b &= \langle 2\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|a^*, \tfrac{1}{2}(\overline{x}_1 + \overline{x}_2)\rangle \\
&= \langle 2W^{-1}(\overline{x}_1 - \overline{x}_2), \tfrac{1}{2}(\overline{x}_1 + \overline{x}_2)\rangle \\
&= -\langle W^{-1}\overline{x}_2, \overline{x}_1\rangle - \langle W^{-1}\overline{x}_2, \overline{x}_2\rangle + \langle W^{-1}\overline{x}_1, \overline{x}_1\rangle + \langle W^{-1}\overline{x}_1, \overline{x}_2\rangle \\
&= \overline{x}_1^t W^{-1}\overline{x}_1 - \overline{x}_2^t W^{-1}\overline{x}_2.
\end{aligned}$$

Es folgt:

$$\widehat{d}(x) = 2\left(-2\|W^{-1}(\overline{x}_1 - \overline{x}_2)\|(\langle a^*, x\rangle + b) - \ln(\frac{w_1}{w_2})\right).$$

Aus dieser Formel und ihrer Herleitung ergibt sich folgender Sachverhalt:

SATZ 5.4: *Es liege das in den Punkten 1 bis 11 von Abschnitt 2.2 formulierte Klassifikationsszenario für zwei Klassen ($r = 2$) vor, wobei die Merkmale $X_1, \ldots, X_p$ jeweils $\mathbb{R}$ als Wertebereich besitzen. Weiter seien die beiden Zufallsvariablen $Z_k := (X_1, \ldots, X_p)|_{\Omega_k}$, $k \in \{1, 2\}$, normalverteilt mit Erwartungswert μ_k und identischen Kovarianzmatrizen $\Sigma_1 = \Sigma_2$. Dann gilt:*

1. *Die Funktion $\widehat{d}(x) = \langle \lambda a^*, x\rangle - \lambda b - \ln(\frac{w_1}{w_2})$, wobei a^* und b die Koeffizienten (29) und (30) von Fishers affiner Diskriminanzfunktion, $\lambda = \|W^{-1}(\overline{x}_2 - \overline{x}_1)\|$ und w_1, w_2 Gewichte für die klassenweisen Trefferwahrscheinlichkeiten sind, ist ein Schätzer für die definierende Gleichung (20) des Bayes-Klassifikators aus Satz 3.7.*

2. *Der durch Fishers Diskriminanzfunktion gelieferte Klassifikator (25) ist ein Schätzer für den Bayes-Klassifikator, falls $w_1 = w_2$ gilt, also zum Beispiel bei identischen a-priori-Wahrscheinlichkeiten $p_1 = p_2$ und ohne zusätzliche Gewichte auf den Trefferwahrscheinlichkeiten.*

Der Mehrklassenfall

Im Folgenden soll der Fall von $r \geq 3$ Klassen soweit entwickelt werden, dass ein mit numerischen Methoden lösbares Optimierungsproblem vorliegt. Explizite Formeln für die Diskriminanzfunktionen wie im Fall $r = 2$ werden nicht angegeben.

Es seien also

$$\Lambda = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3 \cup \ldots \cup \Lambda_r$$

die Zerlegung der vorliegenden Stichprobe in r Klassen und

$$(x^{(i)}, y^{(i)}) := (X(\omega_i), Y(\omega_i)),\ i \in \{1, \ldots, n\}$$

die Merkmalsausprägungen der $\omega_i \in \Lambda$. Wie im Zweiklassenfall definiert man die zu den Klassen gehörenden Punktfamilien

$$C_j := (x^{(i)})_{i:y^{(i)}=j}$$

in $\mathbb{R}^p$. Fisher machte nun den folgenden Ansatz für die lineare Diskriminanzanalyse: Man ermittle eine lineare Abbildung

$$\pi : \mathbb{R}^p \to \mathbb{R}^q,\ x \mapsto Ax, \quad A \in \mathbb{R}^{q \times p}, \tag{35}$$

für welche die Bildfamilien $\pi(C_1), \ldots, \pi(C_r)$ »gut getrennt« im Raum $\mathbb{R}^q$ liegen. Nehmen wir an eine solche Abbildung π also Matrix A wurde bestimmt. Ist dann

$$\overline{x}_j = \frac{1}{|C_j|} \sum_{y^i=j} x^{(i)}$$

der geometrische Schwerpunkt von C_j, so ist wegen der Linearität von π das Bild $\pi(\overline{x}_j)$ der geometrische Schwerpunkt von $\pi(C_j)$ und es wird nach folgender Regel klassifiziert:

Ein $x \in S_X$ wird genau dann der Klasse k zugeordnet, wenn

$$\|\pi(x) - \pi(\overline{x}_k)\| = \min(\|\pi(x) - \pi(\overline{x}_j)\| : j \in \{1, \ldots, r\}) \tag{36}$$

gilt.

Hierbei kann man im Prinzip jede Norm $\|\cdot\|$ auf $\mathbb{R}^p$ nutzen. Die Diskriminanzfunktionen sind also

$$d_j(x) = -\|\pi(x) - \pi(\overline{x}_j)\|. \tag{37}$$

Um die Matrix A zu ermitteln ist zu definieren, was »gut getrennt« bedeuten soll. Wie im 2-Klassen-Fall benutzt man hierfür empirische Kovarianzmatrizen (auch *Streuungsmatrix*; engl.: *scatter matrix*). Die folgende Darstellung erfolgt für die Familien C_j im Originalraum $\mathbb{R}^p$, obwohl wir später die mittels der Abbildung π transformierten Familien betrachten werden.

Mit Hilfe der klassenweisen empirischen Kovarianzmatrizen

$$W_j := \frac{1}{|C_j|} \sum_{y^{(i)}=j} (x^{(i)} - \overline{x}_j)(x^{(i)} - \overline{x}_j)^t \tag{38}$$

definiert man die *Innerklassenstreumatrix* (engl.: *within-class scatter matrix*)

$$W := \sum_{j=1}^{r} \frac{|C_j|}{|\Lambda|} W_j. \tag{39}$$

Die *Zwischenklassenstreumatrix* (engl.: *between-class scatter matrix*) ist als

$$B := \sum_{j=1}^{r} \frac{|C_j|}{|\Lambda|} (\overline{x}_j - \overline{x})(\overline{x}_j - \overline{x})^t \tag{40}$$

definiert. Es gilt dann

$$W + B = \frac{1}{|\Lambda|} \sum_{i=1}^{n} (x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^t \tag{41}$$

mit

$$\overline{x} = \frac{1}{|\Lambda|} \sum_{i=1}^{n} x^{(i)},$$

denn aus

$$\begin{aligned}
(x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^t &= (x^{(i)} - \overline{x}_j + \overline{x}_j - \overline{x})(x^{(i)} - \overline{x}_j + \overline{x}_j - \overline{x})^t \\
&= (x^{(i)} - \overline{x}_j)(x^{(i)} - \overline{x}_j)^t + (\overline{x}_j - \overline{x})(\overline{x}_j - \overline{x})^t \\
&+ (x^{(i)} - \overline{x}_j)(\overline{x}_j - \overline{x})^t + (\overline{x}_j - \overline{x})(x^{(i)} - \overline{x}_j)^t
\end{aligned}$$

folgt

$$\begin{aligned}
\sum_{i=1}^{n} (x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^t &= \sum_{j=1}^{r} \sum_{y^{(i)}=j} (x^{(i)} - \overline{x}_j)(x^{(i)} - \overline{x}_j)^t + \sum_{j=1}^{r} |C_j|(\overline{x}_j - \overline{x})(\overline{x}_j - \overline{x})^t \\
&+ \sum_{j=1}^{r} \sum_{y^{(i)}=j} (x^{(i)} - \overline{x}_j)(\overline{x}_j - \overline{x})^t + (\overline{x}_j - \overline{x})(x^{(i)} - \overline{x}_j)^t \\
&= \sum_{j=1}^{r} |C_j| W_j + |\Lambda| B \\
&+ \sum_{j=1}^{r} (|C_j|\overline{x}_j - |C_j|\overline{x}_j)(\overline{x}_j - \overline{x})^t + (\overline{x}_j - \overline{x})(|C_j|\overline{x}_j - |C_j|\overline{x}_j)^t \\
&= |\Lambda| W + |\Lambda| B|,
\end{aligned}$$

also die Behauptung.

FESTSTELLUNG 5.5: *Die Matrizen W und B sind symmetrisch und positiv semidefinit. Die Matrix W ist genau dann positiv definit, wenn es einen Untervektorraum $U \neq \mathbb{R}^p$ gibt, sodass $C_j \subset \overline{x}_j + U$ für jede Familie C_j gilt.*

BEWEIS: Summen symmetrischer, positiv semidefiniter Matrizen sind symmetrisch und positiv semidefinit; dasselbe gilt für nicht-negative Vielfache solcher Matrizen. Daher genügt es wegen der speziellen Form von W und B die Symmetrie und positive Semidefinitheit für Matrizen der Form xx^t mit $x \in \mathbb{R}^p$ zu beweisen. Die Symmetrie ist dann offensichtlich. Hinsichtlich der positiven Semidefinitheit gilt für beliebiges $z \in \mathbb{R}^p$:

$$z^t(xx^t)z = (z^t x)(x^t z) = \langle z, x\rangle^2 \geq 0. \tag{42}$$

Um die verbleibende Behauptung zu beweisen beginnt man mit der Beobachtung, dass für $z \in \mathbb{R}^p \setminus 0$ die Ungleichung $z^t W z > 0$ genau dann gilt, wenn es ein $j \in \{1, \ldots, r\}$ und ein Sample $x^{(i)} \in C_j$ mit

$$z^t(x^{(i)} - \overline{x}_j)(x^{(i)} - \overline{x}_j)^t z = \langle z, x^{(i)} - \overline{x}_j\rangle > 0$$

gibt. Ist dies nicht der Fall, so folgt

$$x^{(i)} - \overline{x}_j \perp z$$

für alle Samples in C_j, das heißt $C_j \subset \overline{x}_j + U$, wobei U der Orthogonalraum der Ursprungsgeraden $\mathbb{R}z$ ist. □

Man kann sich bei den weiteren Betrachtungen auf den Fall einer positiv definiten Matrix W beschränken: Andernfalls sei U der Orthogonalraum der Ursprungsgeraden $g := \mathbb{R}z$ wie im Beweis von Feststellung 5.5 und $\pi_g : \mathbb{R}^p \to g$ die orthogonale Projektion auf g. Dann gilt

$$\{\pi_g(C_j) : j \in \{1, \ldots, r\}\} = \{z_1, \ldots, z_s\}$$

mit gewissen $z_i \in g$ und $s \leq r$.

Für jedes $j^* \in \{1, \ldots, s\}$ sei $V_{j^*} \subseteq U$ der von den Vektoren

$$x^{(i)} - \overline{x}_j, \quad x^{(i)} \in C_j, \pi_g(C_j) = z_{j^*},$$

erzeugte Untervektorraum von $\mathbb{R}^p$.

Mit Hilfe der orthogonalen Projektionen $\pi_{V_{j^*}} : \mathbb{R}^p \to V_{j^*}$ definiert man die Familien

$$E_{j^*} := \bigcup_{\pi_g(C_j)=z_{j^*}} \pi_{V_{j^*}}(C_j). \tag{43}$$

in V_{j^*}.

Nun kann man das ursprüngliche Klassifikationsproblem in folgender Weise mittels linearer Diskriminanzfunktionen lösen:

A. Ordne ein gegebenes Sample $x \in \mathbb{R}^p$ zunächst derjenigen Klasse $k^* \in \{1, \ldots, s\}$ zu, für die $\|\pi_g(x) - z_{j^*}\|$ minimal ist.

B. Klassifiziere dann $\pi_{V_{j^*}}(x)$ gemäß dem noch zu beschreibenden Verfahren angewandt auf das Klassifikationsproblem (43) in dem Unterraum V_{j^*}.

Man beachte, dass die im Schritt B auftretende Innerklassenstreumatrix nach Definition von V_{j^*} positiv definit ist.

Je nach Lage der Punktfamilien C_j können verschiedene Extremfälle auftreten:

- $s = 1$, das heißt es genügt die Betrachtung eines Untervektorraums U.

  Dieser Fall tritt wohl am häufigsten auf. Anschaulich liegen die Datenpunkt $x^{(i)} \in X(\lambda)$ einfach »gut verteilt« in einem echten affinen Unterraum von $\mathbb{R}^p$.

- $s = r$, das heißt der Schritt B ist nicht erforderlich, da jede Punktklasse C_j in einem »eigenen« affinen Unterraum liegt, der bereits mit Schritt A identifiziert wird.

Ab jetzt sei also die Matrix W positiv definit.

Nach Definition gilt für die Spur von W

$$\operatorname{tr}(W) = \frac{1}{|\Lambda|} \sum_{j=1}^{r} \sigma_j^2,$$

wobei σ_j^2 die empirische Varianz in der Punktfamilie C_j ist. Weiter ist

$$\operatorname{tr}(W + B) = \frac{1}{|\Lambda|} \sigma^2,$$

wobei σ^2 die empirische Varianz in der Punktfamilie $X(\Lambda)$ ist. Damit ist klar, dass die Punktfamilien »gut getrennt« liegen, wenn die Kennzahl

$$J := \frac{\mathrm{tr}\,(W+B)}{\mathrm{tr}\,(W)} = \frac{\sigma^2}{\sum_{j=1}^{r} \sigma_j^2} \tag{44}$$

einen »großen« Wert annimmt.

Wir wenden die zuletzt gewonnen Erkenntnisse auf die Bildfamilien $\pi(C_1), \ldots, \pi(C_r)$ zu einer durch eine Matrix $A \in \mathbb{R}^{q\times p}$ gegebenen linearen Abbildung (35) an: Für die empirische Kovarianzmatrix der Punktfamilie $\pi(C_j)$ erhält man

$$\begin{aligned} W_{j,\pi} &= \tfrac{1}{|\pi(C_j)|} \sum_{y^{(i)}=j} (\pi(x^{(i)}) - \pi(\overline{x}_j))(\pi(x^{(i)}) - \pi(\overline{x}_j))^t \\ &= \tfrac{1}{|C_j|} \sum_{y^{(i)}=j} A(x^{(i)} - \overline{x}_j)(A(x^{(i)} - \overline{x}_j))^t \\ &= AW_jA^t. \end{aligned}$$

Analog ergibt sich für die Zwischenklassenstreumatrix der Bildfamilien $\pi(C_1), \ldots, \pi(C_r)$ die Gleichung

$$B_\pi = ABA^t.$$

Insgesamt haben wir bewiesen:
Die Matrix $A^* \in \mathbb{R}^{q\times p}$ definiert eine in Bezug auf den Klassifikator (36) optimale lineare Abbildung (35), wenn sie das Optimierungproblem

$$\mathrm{argmax}\left(\frac{\mathrm{tr}\,(A(W+B)A^t)}{\mathrm{tr}\,(AWA^t)} : A \in \mathbb{R}^{q\times p}\right)$$

löst.
In Kapitel 5 von [TK] wird gezeigt, dass man anstelle der Spur im obigen Optimierungsproblem auch die Determinante benutzen kann. In diesem Fall lässt es sich in ein verallgemeinertes Eigenwertproblem umformulieren, was numerisch gut behandelbar ist.

6 Klassifikationsbäume

Die bisher betrachteten Modellräume $\mathbf{F} \subseteq \mathrm{M}(S_X, \{1, 2, \ldots, r\})$, in denen nach Klassifikatoren gesucht wird, besitzen einen Nachteil: Für ein typisches Element $f \in \mathbf{F}$ lässt sich nicht direkt nachvollziehen, welche Merkmalsausprägungen eines Objekts $\omega \in \Omega$ mit welcher »Intensität« zum Klassifikationsergebnis $f(X(\omega))$beitragen. Der Klassifikator f erscheint als »Black Box«, deren Inneres dem Anwender verborgen bleibt. In diesem Abschnitt wird mit den Klassifikationsbäumen ein Typ von Klassifikatoren vorgestellt, die im Hinblick auf die Interpretierbarkeit des Klassifikationsergebnisses das genaue Gegenteil von Black Boxes sind. Für diesen Vorteil zahlt man allerdings einen Preis: Typischerweise liegen die Trefferquoten von Klassifikationsbäumen niedriger als die vergleichbarer Black-Box-Klassifikatoren.

6.1 Einfach interpretierbare Klassifikatoren

Es sei $S_X = S_1 \times \ldots \times S_p$ der Merkmalsraum zu den Inputs $X = (X_1, \ldots, X_p)$ eines Klassifikationsproblems. Die Merkmale X_i müssen nicht alle denselben Typ besitzen. Wir betrachten folgende Bedingungen, die von den Ausprägungen eines Merkmals X_i abhängig von dessen Typ erfüllt sein können:

I: $X_i \in T$, $T \subseteq S_i$ eine endlich Menge (beliebiger Typ),

II: $X_i \leq C$, $C \in S_i$ (ordinaler Typ),

III: $X_i \geq C$, $C \in S_i$ (ordinaler Typ).

Definition 6.1: *Ein Klassifikator $f : S_X \to \{1, 2, \ldots, r\}$ heißt einfach interpretierbar, falls es eine endliche Menge $\mathcal{B}$ von Bedingungen der Formen I bis III sowie deren logischer Negationen gibt, sodass Folgendes gilt: Für jedes $k \in \{1, 2, \ldots, r\}$ gibt es Folgen $(B_{1,1}, \ldots, B_{1,s_1}), \ldots, (B_{m_k,1}, \ldots, B_{m_k,s_{m_k}})$ in $\mathcal{B}$ mit der Eigenschaft*

$$f(x_1, \ldots, x_p) = k \;\Leftrightarrow\; \begin{cases} x_1, \ldots, x_p & \textit{erfüllen} & B_{1,1} \wedge \ldots \wedge B_{1,s_1} \\ & \vee & \\ x_1, \ldots, x_p & \textit{erfüllen} & B_{2,1} \wedge \ldots \wedge B_{2,s_2} \\ & \vee & \\ & \vdots & \\ & \vee & \\ x_1, \ldots, x_p & \textit{erfüllen} & B_{m_k,1} \wedge \ldots \wedge B_{m_k,s_{m_k}}. \end{cases}$$

Jeder Funktionswert $f(x_1, \ldots, x_p)$ eines einfach interpretierbaren Klassifikators kann also bestimmt werden, indem man endlich viele logische Verknüpfungen von Bedingungen der Formen I bis III und deren Negationen für die Ausprägungen $x_1, \ldots, x_p$ testet. Wesentlich ist dabei, dass jede der Bedingungen der Form I bis III jeweils nur eine einzelne Ausprägung betrifft und daher unmittelbar interpretierbar ist.

Die Bestimmung der Bedingungsmenge $\mathcal{B}$ und der logischen Verknüpfungen aus einer Stichprobe Λ ist ohne weitere Voraussetzungen komplex und sehr rechenaufwendig. Daher betrachtet man in der Praxis nur solche einfach interpretierbaren Klassifikatoren, die durch eine »baumartig« strukturierte Menge von Bedingungen definiert werden können. Um dies zu präzisieren werden einige Begriffe aus der Graphentheorie benötigt:

DEFINITION 6.2: *Ein endlicher gerichteter Graph G ist ein Tripel (E, K, v) bestehend aus zwei endlichen Mengen E und K sowie einer Abbildung $v : K \to E \times E$. Die Elemente von E heißen Ecken, die von K Kanten von G. Gilt $v(k) = (e_1, e_2)$, so bezeichnet man k als die e_1 und e_2 verbindende Kante und nennt e_1 den Anfangs- und e_2 den Endpunkt der Kante.*

Im Weiteren spielen Verbindungen zweier Ecken eines Graphen eine besondere Rolle:

DEFINITION 6.3: *Es sei $G = (E, K, v)$ ein endlicher gerichteter Graph. Ein die Ecken e und e' verbindender Kantenzug ist eine Folge $k_1, \ldots, k_\ell$ in K mit den Eigenschaften: $v(k_1) = (e, \overline{e})$, $v(k_\ell) = (\tilde{e}, e')$ und der Endpunkt jeder Kante k_i stimmt mit dem Anfangspunkt der Kante k_{i+1} überein.*

Wir benötigen sehr spezielle gerichtete Graphen:

DEFINITION 6.4: *Ein Baum T ist ein endlicher gerichteter Graph (E, K, v) mit folgenden Eigenschaften:*

1. *Zusammenhang: Zu je zwei Ecken $e \neq e'$ gibt es eine Ecke e'' derart, dass e'' und e sowie e'' und e' einen verbindenden Kantenzug besitzen.*
2. *Keine Schleifen: Sind die Ecken e und e' durch einen Kantenzug verbindbar, so ist dieser eindeutig bestimmt.*

Ein Baum T heißt binär, falls für jede Ecke $e \in E$ die Bedingung

$$|\{k \in K : v(k) = (e, \overline{e})\}| = 2$$

erfüllt ist.

Bemerkungen:

1. Um im Bild zu bleiben, kann man die Ecken eines Baums T auch besser als »Verzweigungen« von T bezeichnen.

2. Es existiert eine eindeutig bestimmte Ecke e_0 (»Stamm«), die kein Endpunkt einer Kante ist.

3. Es existieren endlich viele Ecken $b_1, \ldots, b_m$ (»Blätter«), die keine Anfangspunkte einer Kante sind.

Definition 6.5: *Ein binärer Baum $T = (E, K, v)$ heißt gelabelt mit der Labelabbildung $\lambda : K \to \{-1, +1\}$, falls für je zwei Kanten $k \neq k'$ mit demselben Anfangspunkt $\lambda(k) \neq \lambda(k')$ gilt.*

Man beachte, dass in einem binären Baum von jeder Ecke $e \in E$ genau zwei Kanten ausgehen.

Wir können nun das ursprüngliche Ziel angehen und einen einfach interpretierbaren Klassifikator $f : S_X \to \{1, \ldots, r\}$ mit Hilfe eines Baums definieren: Hierzu sei $T := (E, K, v, \lambda)$ ein gelabelter binärer Baum, dessen Eckenmenge die Gestalt

$$E = \mathcal{B} \cup L$$

besitzt, wobei die Elemente von $\mathcal{B}$ Bedingungen der Form I, II oder II und die Elemente von L genau die Blätter des Baums sind. Weiter sei $B_0 \in E$ der Stamm von T und

$$\kappa : L \to \{1, \ldots, r\}$$

eine beliebige Abbildung. Der durch T und κ definierte Klassifikator

$$f_{T,\kappa} : S_X \to \{1, \ldots, r\}$$

ist dann wie folgt gegeben:

Für $x = (x_1, \ldots, x_p) \in S_X$ gilt $f_{T,\kappa}(x) = k$ genau dann, wenn es in T einen Kantenzug $B_0, B_1, \ldots, B_s, e$ vom Stamm B_0 zu einem Blatt $\ell \in L$ mit $\kappa(\ell) = k$ gibt, für den die Merkmalsausprägungen $x_1, \ldots, x_p$ die Bedingung

$$C_0 \wedge C_1 \wedge \ldots \wedge C_s$$

erfüllen, wobei C_i durch

$$C_i := \begin{cases} B_i & \text{falls} \quad \lambda(k) = 1 \text{ für } v(k) = (B_i, B_{i+1}) \\ \neg B_i & \text{falls} \quad \lambda(k) = -1 \text{ für } v(k) = (B_i, B_{i+1}) \end{cases}$$

gegeben ist.
Ihrer Konstruktion entsprechend können Klassifikatoren von diesem Typ in Baumform visualisiert werden.

Beispiel 6.6: Wir betrachten zwei reellwertige Merkmale $X_i : \Omega \to \mathbb{R}$ und eine Aufteilung von Ω in drei Klassen $Y : \Omega \to \{1, 2, 3\}$.

Es gelte $T = (E, K, v, \lambda)$ mit folgenden Daten – siehe Abbildung 29. Der Baum erscheint hier wie in der Literatur weitverbreitet auf den Kopf gestellt:

- $E = \{X_1 \geq \pi, X_2 \geq 1, X_2 \leq 0, X_1 \leq 3\} \cup \{A, B, C, D, E\}$,
- $K = \{(X_1 \geq \pi, X_2 \geq 1), (X_1 \geq \pi, X_2 \leq 0), (X_2 \leq 0, X_1 \leq 3),$
  $(X_2 \geq 1, A), (X_2 \geq 1, B), (X_2 \leq 0, C), (X_1 \leq 3, D), (X_1 \leq 3, E)\}$,
- $\lambda((X_1 \geq \pi, X_2 \geq 1)) = 1$, $\lambda((X_1 \geq \pi, X_2 \leq 0) = -1$,
  $\lambda((X_2 \leq 0, X_1 \leq 3) = -1$, $\lambda((X_2 \geq 1, A) = -1$, $\lambda((X_2 \geq 1, B) = 1$,
  $\lambda((X_2 \leq 0, C) = 1$, $\lambda((X_1 \leq 3, D) = -1$, $\lambda((X_1 \leq 3, E) = 1$.

  In der Abbildung wird der Wert -1 als »Nein« und entsprechend der Wert 1 als »Ja« interpretiert.
- $\kappa(A) = 1$, $\kappa(B) = 2$, $\kappa(C) = 2$, $\kappa(D) = 3$, $\kappa(E) = 1$.

Die Kantenabbildung v ergibt sich bei Notation der Kanten als Paare von Ecken von selbst. $\diamond$

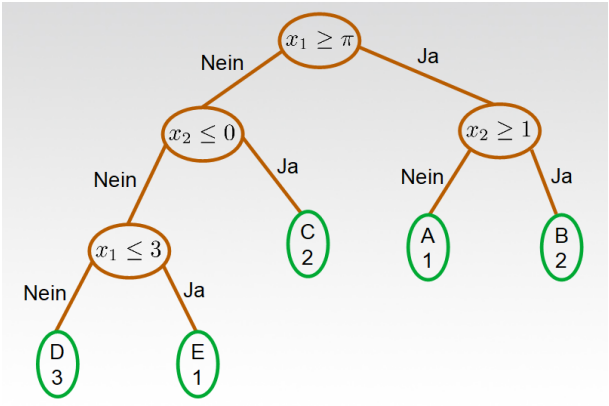


Abbildung 29: Baumartiger Klassifikator

6.2 Binäre Klassifikationsbäume

Klassifikatoren von dem im vorigen Abschnitt eingeführten Typ $f_{T,\kappa}$ beziehungsweise ihre Visualisierung als Baum werden als *binäre Klassifikationsbäume* bezeichnet. Sie wurden von dem Statistiker Leo Breiman[4] etwa im Jahr 1976 eingeführt, während er als Consultant für die Firma »William Meisel's division of Technology Services Corporation« arbeitete. Breiman schrieb später das Buch [Bre] über Klassifikations- und Regressionsbäume, das als grundlegend für das Gebiet gilt.

Schätzung/Training von binären Klassifikationsbäumen
Im Folgenden soll ein Verfahren angegeben werden um einen binären Klassifikationsbaum $f_{T,\kappa}$, $T = (E, K, v, \lambda)$, mit möglichst hoher Trefferwahrscheinlichkeit aus einer gegebenen Stichprobe Λ zu ermitteln. Für das Verständnis des Verfahrens ist es durchgängig nützlich sich den Klassifikator $f_{T,\kappa}$ als Baum wie in Abbildung 29 vorzustellen. Dieser Vorstellung entsprechend wird der gesuchte Klassifikator schrittweise mit dem Stamm beginnend aufgebaut, wobei in jedem Schritt die folgenden grundlegenden Fragen beantwortet werden müssen:

- Soll an das Ende einer bereits vorhandenen Kante (eines »Astes«) eine Verzweigung oder ein Blatt angefügt werden?

[4]Leo Breiman, amerikanischer Statistiker, 1928 – 2005

- Im Falle einer Verzweigung: Welche Bedingung vom Typ I, II oder III soll dort vorliegen?
- Im Falle eines Blattes: Welche Klasse soll diesem Blatt zugeordnet werden?

Die Antwort wird (natürlich) auf der Grundlage der Stichprobe Λ gegeben. Die dabei verfolgte Grundidee ist in Abbildung 30 dargestellt:

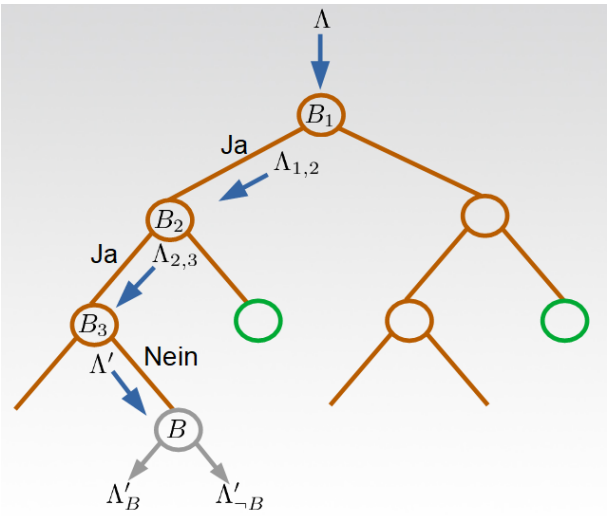


Abbildung 30: Rekursive Erzeugung eines Klassifikationsbaums

Wir nehmen an, dass an der grau markierten Stelle des Baums entschieden werden soll, ob eine Verzweigung oder ein Blatt angehängt wird; der Stamm des Baumes sowie vier Verzweigungen und zwei Blätter wurden bereits festgelegt. Man betrachtet nun diejenige Teilfamilie Λ' der Stichprobe Λ, deren Samples alle Bedingungen längs des eindeutig bestimmten Pfades durch den bereits bestimmten Teil des Klassifikationsbaumes zu der in Rede stehenden Ecke (siehe Punkt 3 der Definition 6.4) erfüllen. Im Beispiel sind das die Bedingungen B_1, B_2 und $\neg B_3$. Im Folgenden wird Λ' auch als *(in die Ecke) einlaufende Restfamilie* bezeichnet.

Jede Bedingung B liefert eine Partition

$$\Lambda' = \Lambda'_B \cup \Lambda'_{\neg B} \tag{45}$$

von Λ' in diejenigen Samples, die B erfüllen, und diejenigen, die dies nicht tun. Die Idee zur Wahl von B besteht nun darin dies so zu tun, dass die verschiedenen Klassen $1, 2, \ldots, r$ in den Familien Λ'_B und $\Lambda'_{\neg B}$ »weniger stark gemischt« vorkommen als in der einlaufenden Restfamilie Λ'. Sollte dies nicht möglich sein, wird an der Ecke ein Blatt angehängt. Um den Begriff »Mischung« zu präzissieren und zu quantifizieren benutzt man sogenannte Unreinheitsfunktionen:

Definition 6.7: *Ein Unreinheitsfunktion ist eine Funktion*

$$\phi : \{(q_1, \ldots, q_r) \in (\mathbb{R}^{\geq 0})^r : q_1 + \ldots + q_r = 1\} \to \mathbb{R}$$

mit den Eigenschaften:

1. *ϕ ist symmetrisch,*
2. *ϕ besitzt ein einziges Maximum und dieses liegt bei $(\frac{1}{r}, \ldots, \frac{1}{r})$,*
3. *ϕ besitzt genau r Minima und eines liegt bei $(1, 0, \ldots, 0)$.*

Die »Gemischtheit« oder »Unreinheit« einer Teilfamilie $\Lambda' \subseteq \Lambda$ lässt sich nun wie folgt quantitativ erfassen:

Definition 6.8: *Es sei ϕ eine Unreinheitsfunktion. Die ϕ-Unreinheit einer Teilfamilie Λ' einer in r Klassen zerfallenden Stichprobe*

$$\Lambda = \Lambda_1 \cup \Lambda_2 \cup \ldots \cup \Lambda_r$$

ist definiert als

$$i_\phi(\Lambda') := \phi(\frac{|\Lambda' \cap \Lambda_1|}{|\Lambda'|}, \ldots, \frac{|\Lambda' \cap \Lambda_r|}{|\Lambda'|}).$$

Bemerkung: Nach Definition einer Unreinheitsfunktion nimmt $i_\phi(\Lambda')$ denn maximalen Wert an, wenn in Λ' alle Klassen gleich häufig vertreten sind. Der minimale Wert wird angenommen, wenn alle Samples in Λ' zu derselben Klasse gehören.

Die folgenden beiden Unreinheitsfunktionen werden häufig verwendet:

- die *Entropie* $\phi(q_1, \ldots, q_r) = -\sum_{k=1}^{r} q_k \log(q_k)$,
- der *Gini*[5]*-Index* $\phi(q_1, \ldots, q_r) = 1 - \sum_{k=1}^{r} q_k^2$.

Der Beweis dafür, dass es sich bei beiden Funktionen um Unreinheitsfunktionen handelt, wird dem Leser überlassen.

Wir können nun den durch eine Bedingung B in einem binären Klassifikationsbaum induzierten *Reinheitsgewinn* definieren:

DEFINITION 6.9: *Es sei B eine Bedingung vom Typ I, II oder III und Λ' eine Teilfamilie einer in r Klassen zerfallenden Stichprobe Λ. Dann ist der durch B für Λ' induzierte ϕ-Reinheitsgewinn durch*

$$\Delta i_\phi(B, \Lambda') := i_\phi(\Lambda') - (\frac{|\Lambda'_B|}{|\Lambda'|} i_\phi(\Lambda'_B) + \frac{|\Lambda'_{\neg B}|}{|\Lambda'|} i_\phi(\Lambda'_{\neg B}))$$

definiert – siehe (45).

BEMERKUNG: Die Gewichtung der ϕ-Unreinheiten von Λ'_B und $\Lambda'_{\neg B}$ mit deren relativen Größen ist notwendig um Overfitting zu vermeiden: Eine Teilmenge mit einem Element besitzt stets die ϕ-Unreinheit 0, womit Partitionen der Form $\Lambda' = \{x\} \cup \Lambda' \setminus \{x\}$ als akzeptabel eingestuft würden, sobald die ϕ-Unreinheit von Λ' klein genug ist. Tritt diese Situation bei der Schätzung von $f_{T,\kappa}$ häufig auf, so erzeugt man einen Klassifikationsbaum mit vielen Verzweigungen, die nur der Klassifikation sehr weniger Samples dienen – Overfitting!

Der Basisalgorithmus zur Schätzung eines binären Klassifikationsbaums kann nun formuliert werden. Hierbei seien:

- $X_1, \ldots, X_p$: die Inputs eines Klassifikationsproblems mit r Klassen;
- Λ: eine Stichprobe zum obigen Klassifikationsproblem;
- ϕ: eine vom Anwender gewählte Unreinheitsfunktion;
- $\mathcal{B}_\Lambda$ die Menge aller Bedingungen vom Typ I, II und III, die folgenden Forderungen genügen:

[5]Corrado Gini, italienischer Statistiker und Soziologe, 1884 – 1965.

1. $T \subseteq X_i(\Lambda)$ für jede Bedingung vom Typ I,
2. $C \in X_i(\Lambda)$ für jede Bedingung vom Typ II und III.

Man beachte, dass es genau $2^{|X_i(\Lambda)|} - 2$ echte Teilmengen T von $X_i(\Lambda)$ gibt und dass die Anzahl relevanter Schranken C gleich $|X_i(\Lambda)| - 1$ ist. Insgesamt ist die Menge $\mathcal{B}_\Lambda$ also endlich.

Algorithmus zur Bestimmung eines binären Klassifikationsbaums

1. **Initialisierung:**

   Wahl eines minimalen Reinheitsgewinns $\Delta_{\min} > 0$.

   Optional: Wahl der maximalen Höhe $h_{\max}$ des Baums T.

   Optional: Wahl der Mindestgröße $n_{\min}$ der in eine Verzweigung einlaufenden Restfamilie (»minimale Verzweigungsgröße«).

   Optional: Wahl der Mindestgröße n_L der in ein Blatt einlaufenden Restfamilie (»minimale Blattgröße«).

   Initialisierung des Baums T: $\mathcal{B} \cup L = E := \emptyset$, $K = \emptyset$.

2. **Suche nach einer potentiellen Verzweigung:** Wähle eine Verzweigung $B_a \in \mathcal{B}$ zu der noch keine zwei Äste $(B_a, B_e) \in K$ existieren.

   Falls es ein solches B_a nicht gibt, gehe zu Schritt 5.

   Optional: Ist die Länge des Pfades vom Stamm B_0 bis B_a gleich $h_{\max}-1$, gehe zu Schritt 4.

   Bestimme die in die zu definierende Ecke e einlaufende Restfamilie Λ'.

   Optional: Falls $|\Lambda'| < n_{\min}$, gehe zu Schritt 4.

   Bestimme eine Bedingung $B \in \mathcal{B}_\Lambda$ mit der Optimalitätseigenschaft

   $$\Delta i_\phi(B, \Lambda') = \max(\Delta i_\phi(B', \Lambda') : B' \in \mathcal{B}_\Lambda). \tag{46}$$

   Optional: Ist $|\Lambda'_B| < n_L$ oder $|\Lambda'_{\neg B}| < n_L$, gehe zu Schritt 4.

   Ist $\Delta i_\phi(B, \Lambda') \geq \Delta_{\min}$, gehe zu Schritt 3 andernfalls zu Schritt 4.

3. **Hinzufügen einer Verzweigung:**

   Setze $\mathcal{B} := \mathcal{B} \cup \{B\}$ und $K := K \cup \{(B_a, B)\}$.

Setze $\lambda((B_a, B)) = 1$, falls für die Bestimmung von Λ' die Bedingung B_a verwendet wurde; setze $\lambda((B_a, B)) = -1$, falls für die Bestimmung von Λ' die Bedingung $\neg B_a$ verwendet wurde.

Gehe zu Schritt 2.

4. **Hinzufügen eines Blattes:**

   Erweitere die Menge der Blätter L und ein weiteres Element ℓ und setze $K := K \cup \{(B_a, \ell)\}$.

   Gehe zu Schritt 2.

5. **Bestimmen von κ:**

   Lege für jedes Blatt $\ell \in L$ den Wert $\kappa(\ell)$ auf folgende Weise fest: Es sei $(B_a, \ell) \in K$ der Ast zum Blatt ℓ und Λ' die Teilfamilie von Λ, deren Samples alle Bedingungen auf dem Pfad vom Stamm B_0 bis B_a erfüllen. Setze

   $$\kappa(\ell) := k,$$

   wenn k die am häufigsten in Λ' vorkommende Klasse ist.

Bemerkungen:

1. Der Basisalgorithmus endet, wenn jedes Element von E entweder zwei ausgehende Äste besitzt oder von der Form (B, ℓ) mit einem Blatt $\ell \in L$ ist. Dies tritt nach endlich vielen Schritten ein, da die Menge potentieller Verzweigungen $\mathcal{B}_\Lambda$ endlich ist und jede Bedingung nach Konstruktionsvorschrift höchstens einmal Verwendung findet.

2. Die Suche nach dem Maximum in (46) kann durch Anwendervorgaben wie zum Beispiel die maximale Größe der Teilmengen T in Bedingungen vom Typ I oder die Angabe, dass für ordinale Merkmale nur die Bedingungen vom Typ II und III verwendet werden sollen, eingeschränkt werden. Dies kann offensichtlich erhebliche Laufzeitvorteile einbringen.

3. Falls die Klasse k im Schritt 5 nicht eindeutig bestimmt ist, kann man aus den Kandidaten einen zufällig wählen.

4. Die im Schritt 5 definierte Abbildung κ muss nicht surjektiv sein. Tritt dieser Fall ein, sollten die verwendeten Abbruchkriterien hinterfragt werden. Möglicherweise lässt die vorliegende Stichprobe aber auch keine

Klassifikation in eine bestimmte Klasse mit akzeptabler Treffsicherheit und Generalisierungsfähigkeit zu.

5. Wie bereits im Beispiel 6.6 angemerkt, ist die konkrete Angabe der Abbildung v nicht notwendig.

   Weiter wird die Angabe der Kantenlabels $\lambda(B, \ell)$, $\ell \in L$, nicht benötigt – siehe Schritt 4.

Beispiel 6.10 (Diabetes-Risiko): Wir betrachten in diesem Beispiel die von dem Data Engineer Mohammed Mustafa auf Kaggle [Kag] zur Verfügung gestellte Datenmenge »Diabetes prediction dataset«. (Das Copyright der Datenmenge liegt bei den Erstellern; als Basis für Datenanalysen darf sie verwendet werden.)

Die Datenmenge umfasst in Bezug auf die Erkrankung Diabetes relevante demografische und medizinische Angaben zu 100 000 Personen. Im hier diskutierten Beispiel werden die folgenden Merkmale verwendet:

- age (Alter): 0.08 – 80 Jahre;
- gender (Geschlecht): Male, Female;
- hypertension (Bluthochdruck): Yes, No;
- heart disease (Herzerkrankung): Yes, No;
- diabetes (Diabetes): Yes, No;
- bmi (Body–Mass-Index): 10.16 – 95.7;
- HbA1c level (Anteil glykiertes Hämoglobin, Maß für den Blutzuckerspiegel der letzten 8 – 12 Wochen): 3.5% – 9%;
- blood glucose level (Blutzuckerspiegel): 80mg/dl – 300mg/dl.

Es stellt sich die Frage, ob man das Vorliegen einer Diabetes-Erkrankung anhand dieser Merkmale (ohne das Merkmal diabetes) vorhersagen kann. Ein entsprechender Klassifikator könnte dann zur Bewertung des Risikos an Diabetes zu erkranken genutzt werden.

In der vorliegenden Datenmenge stammen ca. 9% der Samples von diabetes-kranken Personen. Laut Angaben des Robert-Koch-Instituts [RKI] betrug die Diabetes-Häufigkeit in der deutschen Bevölkerung im Jahr 2010 etwa 9.2%, sodass die Datenmenge auf diesen Zeitpunkt bezogen repräsentativ für die deutsche Bevölkerung ist. Dies bedeutet jedoch auch, dass die Datenmenge in Bezug auf die beiden Klassen »diabetes=No« und »diabetes=Yes« sehr schlecht ausbalanciert ist, was sich negativ auf die Schätzung von Klassifikatoren auswirkt.

In einem ersten Experiment wird ein binärer Klassifikationsbaum f_{T_1,κ_1} auf der Basis einer zufällig aus der Datenmenge gezogenen Stichprobe Λ vom Umfang $|\Lambda| = 10\,000$ erstellt. Die Stichprobe wird klassenweise gezogen (siehe Abschnitt 4), sodass in Λ ebenfalls ca. 9% der Samples von diabetes-kranken Personen stammen. Die Ziehung dieser Stichprobe aus der Gesamtdatenmenge erfolgt aus einem rein technischen Grund: Die verwendete *freie* Version der Software Rapidminer verarbeitet nur Datenmengen mit höchstens 10 000 Samples.

Zur Erzeugung des Baumes werden die folgenden Parameter für den Rapidminer-Operator »Decision Tree« verwendet:

$$\phi = \text{Gini-Index}, \Delta_{\min} = 0.001, h_{\max} = 7, n_{\min} = 20, n_L = 7.$$

Der resultierende Baum ist in Abbildung 31 dargestellt. Die Darstellung unterscheidet sich etwas von derjenigen in Abbildung 29: An den Verzeigungen ist jeweils nur der Name des Merkmals angegeben, das in die dort vorliegende Bedingung B eingeht, während die Bedingung selbst und ihre Negation an den aus B auslaufenden Kanten stehen. Die Liniendicke der Kanten visualisiert die Größe der jeweils in den Endpunkt einlaufenden Restfamilie. Die Angabe in einem Blatt ℓ gibt die dem Blatt zugeordnete Klasse $\kappa(\ell)$ an. Der farbige Balken visualisiert die Anteile von Samples der Klasse »diabetes=Yes« (rot) oder »diabetes=No« (blau) in der einlaufenden Restfamilie.

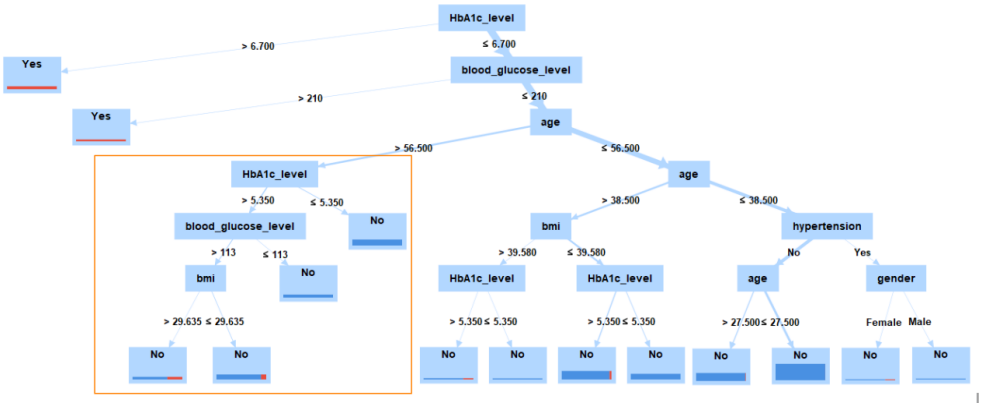


Abbildung 31: Klassifikationsbaum f_{T_1,κ_1} für die Diabetes-Prädiktion

Die mittels 10-facher Kreuzvalidierung geschätzte Trefferwahrscheinlichkeiten des Klassifikationsbaums f_{T_1,κ_1} sind:

$$\begin{aligned}\widehat{P}_{\text{10CV}}(f_{T_1,\kappa_1}) &= 0.9715,\\ \widehat{P}_{\text{10CV}}(f_{T_1,\kappa_1}|\text{diabetes} = \text{No}) &= 1.0,\\ \widehat{P}_{\text{10CV}}(f_{T_1,\kappa_1}|\text{diabetes} = \text{Yes}) &= 0.6791.\end{aligned}$$

Der Klassifikator erkennt also diabetes-kranke Personen deutlich schlechter als Personen ohne Diabetes-Erkrankung. Das Ausmaß der Mängel von f_{T_1,κ_1} erkennt man daran, dass alle Samples mit der Eigenschaft »blood glucose level ≤ 210« als »diabetes=No« klassifiziert werden, obwohl in die Blätter 3,4,5,7 und 11 gezählt von links zusammen 268 Samples mit der Eigenschaft »diabetes=Yes« einlaufen. Der gesamte an dieser Verzweigung hängende Ast könnte also ohne Qualitätseinbuße durch ein Blatt mit der Klassenzuordnung »diabetes=No« ersetzt werden. Abhilfe könnte durch die Betrachtung höherer Bäume geschaffen werden. Da die Blattgrößen von f_{T_1,κ_1} bis auf zwei Ausnahmen größer als 50 sind – die Ausnahmen liegen immer noch oberhalb von 20, wird eine Verkleinerung der zulässigen Blattgröße n_L eher keinen positiven Effekt haben.

In einem zweiten Experiment wird entsprechend ein Klassifikationsbaum f_{T_2,κ_2} mit den Parametern

$$\phi = \text{Gini-Index}, \Delta_{\min} = 0.001, h_{\max} = 10, n_{\min} = 20, n_L = 7$$

erzeugt. Die kreuzvalidierten Trefferquoten sind

$$\begin{aligned}\widehat{P}_{\text{10CV}}(f_{T_2,\kappa_2}) &= 0.9702,\\ \widehat{P}_{\text{10CV}}(f_{T_2,\kappa_2}|\text{diabetes} = \text{No}) &= 0.9962,\\ \widehat{P}_{\text{10CV}}(f_{T_2,\kappa_2}|\text{diabetes} = \text{Yes}) &= 0.7038.\end{aligned}$$

Der leichten Verbesserung in der Klasse »diabetes=Yes« steht eine enorme Erhöhung der Komplexität gegenüber: Die Eckenzahl (ohne Blätter) wächst von 13 auf 43. Im Wesentlichen wird dabei die Erhöhung der Trefferwahrscheinlichkeit in der Klasse »diabetes=Yes« durch den in Abbildung 32 dargestellten Ast bewirkt. Er ergibt sich durch Einfügen zusätzlicher Verzweigungen aus dem in Abbildung 31 orange markierten Ast.

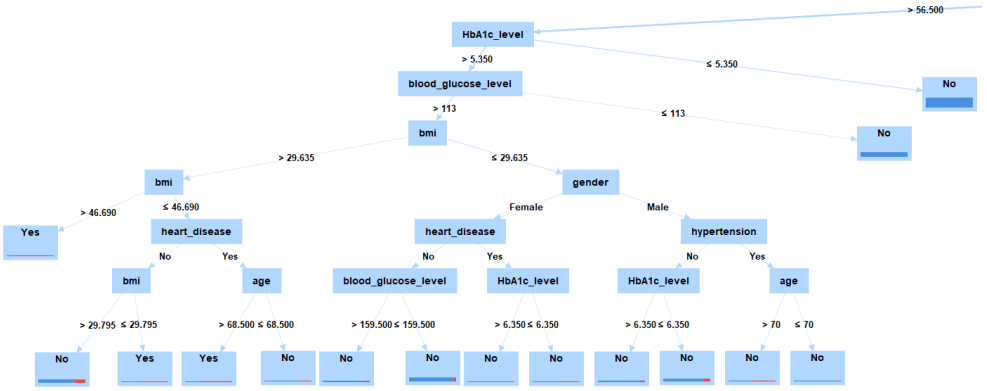


Abbildung 32: Teil des Klassifikationsbaum f_{T_2,κ_2}

Schließlich wird in einem dritten Experiment ein Klassifikationsbaum f_{T_3,κ_3} auf einer aus Λ zufällig erzeugten, balancierten Stichprobe $\overline{\Lambda}$ trainiert. Diese umfasst jeweils 800 Samples von Personen mit und ohne Diabetes-Erkrankung. $\overline{\Lambda}$ enthält damit insbesondere fast alle in Λ vorhandenen Samples von Personen mit Diabetes. Da die Trainingsmenge nun deutlich kleiner ist, werden Bäume geringerer Höhe betrachtet:

$$\phi = \text{Gini-Index}, \Delta_{\min} = 0.001, h_{\max} = 5, n_{\min} = 20, n_L = 7.$$

Abbildung 33 zeigt das Ergebnis für eine solche Zufallsziehung $\overline{\Lambda}$ mit den geschätzten Trefferwahrscheinlichkeiten

$$\begin{aligned} \widehat{P}_{\text{10CV}}(f_{T_3,\kappa_3}) &= 0.8431, \\ \widehat{P}_{\text{10CV}}(f_{T_3,\kappa_3}|\text{diabetes} = \text{No}) &= 0.7625, \\ \widehat{P}_{\text{10CV}}(f_{T_3,\kappa_3}|\text{diabetes} = \text{Yes}) &= 0.9238. \end{aligned}$$

Angesichts der erheblichen Unterschiede zwischen den Schätzungen der Trefferwahrscheinlichkeiten der drei Experimente stellt sich die Frage wie stark die Ergebnisse von der Teilstichprobe $\overline{\Lambda}$ abhängen. Hierzu wurde das Experiment 200 mal mit zufällig gezogenen Teilstichproben $\overline{\Lambda}$ wiederholt. Die Verteilungen der so erhaltenen Trefferwahrscheinlichkeiten sind in den Histogrammen in Abbildung 34 dargestellt. Wie man erkennt liegt hier keine stabile Situation vor: Die klassenweisen Trefferwahrscheinlichkeiten variieren stark. Allerdings sollten besser die 200 Klassifikationsbäume selbst

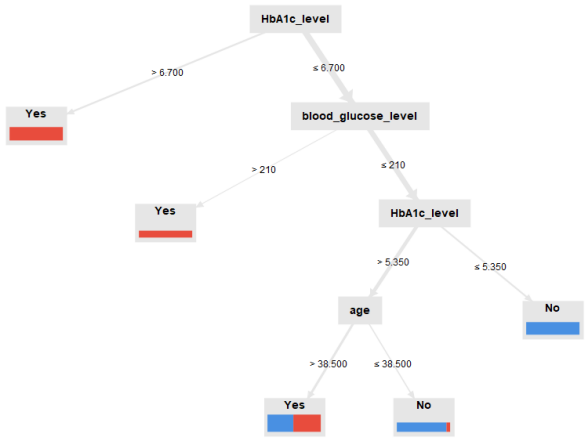


Abbildung 33: Klassifikationsbaum f_{T_3,κ_3}

miteinander verglichen werden. Der Vergleich von Klassifikationsbäumen ist jedoch im Allgemeinen kein einfaches Unterfangen und würde umfangreichere Ausführungen erfordern. ◇

Pruning (Beschneiden)

Ein binärer Klassifikationsbaum kann trotz sorgfältig gewählter Parameter für das Training zu komplex sein: Die Blätter an einzelnen, möglicherweise sogar stark verzweigten Ästen tragen im Verhältnis nur geringfügig zur Güte des Klassifikators bei. Ein Beispiel ist der bereits erwähnte Ast rechts der Verzweigung »age$\leq$ 56.5« in Abbildung 31, der vollständig und ohne Güteverlust durch ein Blatt mit der Klassenzuordnung »diabetes=No« ersetzt werden könnte. Um dieses Phänomen zu behandeln, kann man einen binären Klassifikationsbaum nach der Erstellung systematisch beschneiden: Im Prinzip legt man einen gerade noch akzeptablen Güteverlust ΔT für die totale Trefferwahrscheinlichkeit fest und bestimmt alle Äste des vorliegenden Baums, deren Entfernen und Ersetzen durch ein einzelnes Blatt zu einem Güteverlust kleiner als ΔT führen würde. Aus der erhaltenen Kandidatenmenge wird nach einem vorgegebenen Kriterium gewählt. Beispielsweise könnte man stets den Ast mit der größten Verzweigungszahl auswählen. Generell beruht das Auswahlkriterium auf einem Komplexitätsmaß für Bäume.

Das Pruning soll hier nicht im Detail behandelt werden. Eine ausführliche Darstellung findet sich in Kapitel 10 von [Bre].

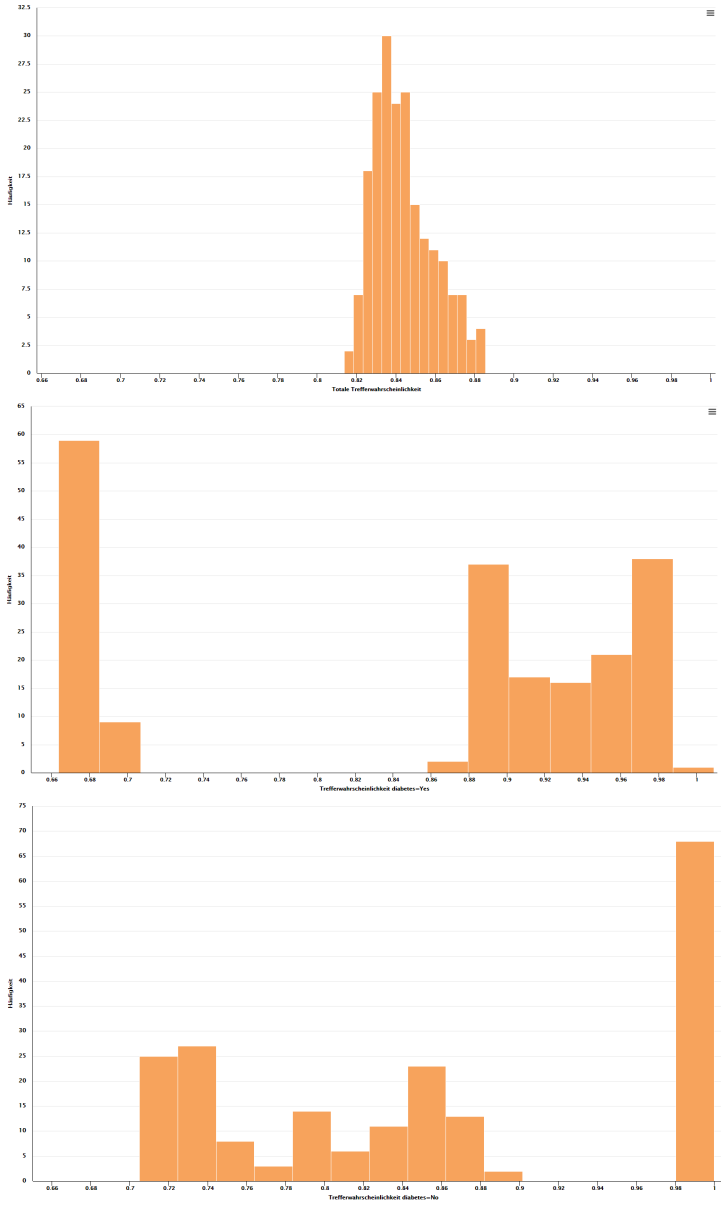


Abbildung 34: Verteilung der Trefferwahrscheinlichkeiten von f_{T_3,κ_3}
Oben: Totale Trefferwahrscheinlichkeiten
Mitte: Trefferwahrscheinlichkeiten diabetes=Yes
Unten: Trefferwahrscheinlichkeiten diabetes=No

Allgemeinere Klassifikationsbäume

Das Konzept des binären Klassifikationsbaums wurde seit seiner Einführung in verschiedene Richtungen erweitert, eine Entwicklung, die von den Begründern der Theorie deutlich kritisiert wurde: Die Erweiterungen führten in der Regel in mehr oder weniger ausgeprägter Weise weg von dem ursprünglichen Motiv einfach interpretierbare Klassifikatoren bestimmen zu wollen.

Im Wesentlichen sind zwei Arten solcher Erweiterungen zu nennen:

- **Nicht-binäre Verzweigungen:** Aus einer Ecke des Klassifikationsbaums dürfen mehr als zwei Kanten auslaufen. Bleibt man bei der für binäre Klassifikationsbäume eingeführten Grundmenge von Bedingungen (Typen I, II oder III), so entspricht eine Verzweigung in mehr als zwei Äste einer logisch komplexeren Bedingung etwa der Form

$$X \leq C_1 \vee X \in T \vee X > C_2.$$

- **Komplexe Bedingungen:** Die Bedingungen vom Typ I, II oder III kann man durch erheblich komplexere ersetzen. So gibt es beispielsweise Klassifikationsbäume, in denen für die Verzweigungsbedingungen die in Abschnitt 5 eingeführten linearen Diskriminanzfunktionen verwendet werden.

7 Merkmalswahl

Das allgemeine Klassifikationsproblem wurde bislang so dargestellt, dass die Merkmale $X_1, \ldots, X_p$ und Y vorgegeben sind, nämlich durch die vorliegende Datenmenge zu einer Stichprobe. In der Praxis ist die Situation komplexer:

Merkmalswahl vor der Datenerhebung: In der empirischen Forschung wie sie in allen Natur- und Sozialwissenschaften betrieben wird, werden die zu betrachtenden Merkmale bereits vor der Datenerfassung festgelegt: Man führt Experimente oder eine Befragung mit einem bestimmten Ziel durch. Dabei werden die im Experiment zu erfassenden Größen beziehungsweise die während der Befragung zu stellenden Fragen vorab festgelegt. Diese Festlegung erfolgt auf der Basis vorliegender Hypothesen unter Einbezug der relevanten Theorie. Selbst wenn man als Datenanalytiker an diesem Vorgang der Merkmalswahl vor der Datenerhebung nicht beteiligt ist, sollte man sich dessen bewusst sein, da Fehler bei der Merkmalswahl die Güte der durch die Datenanalyse möglichen Resultate von vorneherein begrenzen.

Änderung der ursprünglichen Merkmalsbasis: Die Tatsache, dass die Merkmale $X_1, \ldots, X_p$ und Y durch eine vorliegende Datenmenge vorgegeben sind, bedeutet nicht, dass man notwendigerweise direkt mit einer Auswahl dieser Merkmale arbeiten muss. Es kann nützlich sein aus den gegebenen Merkmalen durch Transformationen neue Merkmale zu erzeugen und die ursprünglichen Merkmale entweder durch die neu erzeugten zu ersetzen oder die neu erzeugten zu den ursprünglichen hinzuzunehmen.

Auswahl von Merkmalen aus der Merkmalsbasis: Es kann nützlich oder aufgrund der Datenlage sogar notwendig sein nicht alle durch eine Datenmenge vorgegebenen Merkmale in einer Analyse der Daten wirklich zu benutzen.

Im vorliegenden Abschnitt werden die Themen Merkmalsauswahl und Änderung der Merkmalsbasis addressiert.

7.1 Merkmalsauswahl

In praktischen Klassifikationsproblemen tritt häufig die folgende Frage auf: *Welche aus einer vorliegenden Menge von Merkmalen $X_1, \ldots, X_p$ werden wirklich zur Prognose der Klassenzugehörigkeit Y der betrachteten Objekte benötigt?*

Die Gründe dafür diese Frage zu stellen können ganz unterschiedlich sein:

- Fachlich-wissenschaftliche: Welche Gesetzmäßigkeiten bestimmen die Klassenzugehörigkeit der betrachteten Objekte?
- Pragmatische: Man möchte ein möglichst übersichtliches, gut interpretierbares Modell der vorliegenden Daten erhalten.
- Kontextuelle: Die vorliegenden Daten wurden ursprünglich für einen anderen Zweck als den aktuell betrachteten erhoben. Daher enthalten sie möglicherweise überflüssige Merkmale.
- Data-Mining-spezifische: Die Anzahl der Merkmale ist im Vergleich zum Umfang der Stichprobe (zu) groß, wodurch eine gute Schätzung eines Modells für die Daten in Frage gestellt ist.

Die einfachste Methode um die eingangs gestellte Frage zu beantworten ist das systematische Erzeugen von Klassifikatoren, die jeweils nur Teilmengen der Menge aller zur Verfügung stehenden Merkmale nutzen. Die erzeugten Klassifikatoren werden mittels ihrer geschätzten Trefferwahrscheinlichkeit verglichen und der in dieser Hinsicht beste Klassifikator wird ausgewählt. Dieses Vorgehen ist im allgemeinen *nicht* vereinbar mit der Festlegung *einer einzigen* Klasse $\mathbf{F}$ von Klassifikatoren, innerhalb der man die zu erzeugenden Klassifikatoren wählt, denn der Definitionsbereich eines Klassifikators hängt von dessen Variablenzahl ab. Die skizzierte Vorgehensweise muss präziser gefasst werden: Im Folgenden liege das in den Punkten 1 bis 11 von Abschnitt 2.2 formulierte Klassifikationsszenario mit reellen Merkmalen $X_1, \ldots, X_p$ und im Zwei-Klassen-Fall ($r = 2$) vor. Zu jedem Merkmal X_j liege eine Borelmessbare Funktion

$$g_j : S_j \to \mathbb{R}$$

vor; man kann g_j als eine Transformation des Originalmerkmals betrachten.

Wir betrachten nun Diskriminanzfunktionen der Gestalt

$$d(X_1, \ldots, X_p) = b + \sum_{j=1}^{p} a_j g_j(X_j),\ b, a_1, \ldots, a_p \in \mathbb{R}$$

und den Raum

$$\mathbf{F} := \{\mathrm{sgn}(b + \sum_{j=1}^{p} a_j g_j(X_j)) : b, a_1, \ldots, a_p \in \mathbb{R}\}$$

der zugehörigen Klassifikatoren. Man beachte, dass alle diese Funktionen Borel-messbar sind.

Der so definierte Suchraum erlaubt das »Weglassen« von Merkmalen, indem man den entsprechenden Koeffizienten a_j gleich 0 setzt.

Im Fall $g_j(X_j) = X_j$ liegen in $\mathbf{F}$ genau die von Fisher betrachteten Klassifikatoren aus Basis einer affinen Diskriminanzfunktion.

Best Subset Selection

Es liege die gerade beschriebene Situation vor. Weiter sei eine Methode zur Schätzung der Trefferwahrscheinlichkeit eines Klassifikators *fest vorgegeben*. Zum Beispiel kann eine der Methoden

- Bestimmung der Trefferquote auf der Trainingsmenge,
- Bestimmung der Trefferquote auf einer fest vorgegebenen Testmenge,
- Bestimmung der k-fach kreuzvalidierten Trefferquote bei fest vorgegebener Aufteilung der Trainingsmenge,

genutzt werden.

Schließlich sei eine Methode zur Bestimmung von Klassifikatoren $f \in \mathbf{F}$ auf Basis einer Stichprobe Λ gegeben.

Wie der Name der hier zu beschreibenden Methode bereits vermuten lässt geht man dann wie folgt vor:

1. Wahl einer minimalen und einer maximalen Merkmalsanzahl $1 \leq q_{\min} < q_{\max} \leq p$.
2. Zu jeder Teilmenge $T \subseteq \{X_1, \ldots, X_p\}$ mit mindestens $q_{\min}$ und höchsten $q_{\max}$ Merkmalen: Bestimmung des Klassifikators $f_{T,\Lambda}$ und seiner geschätzten (gewichteten) Trefferwahrscheinlichkeit $\widehat{P}(f_{\Lambda,T}(X_1, \ldots, X_p) = Y)_w$.
3. Wahl derjenigen Teilmenge(n) T_0, für die $\widehat{P}(f_{\Lambda,T_0}(X_1, \ldots, X_p) = Y)_w$ maximal ist/sind.

Bei der Anwendung der Methode ist zu beachten, dass insgesamt

$$\sum_{q=q_{\min}}^{q_{\max}} \binom{p}{q}$$

Klassifikatoren zu betrachten sind. Der Rechenaufwand der Methode ist also gegebenenfalls sehr groß.

Beispiel 7.1: Die folgenden Ausführungen beziehen sich auf die Datenmenge »Chemical Composition of Ceramic« des UCI Machine Learning Repository [UCI]. Die Datenmenge umfasst Angaben zur chemischen Zusammensetzung von 44 Scherben chinesischen Porzellans, wobei für jede Scherbe zwischen der Zusammensetzung der Glasur und der des Porzellankörpers unterschieden wird (Merkmal Part$\in$ {Body,Glaze}). Angegeben sind in beiden Fällen die Masseanteile folgender 17 Oxide:

Na_2O (%), MgO (%), Al_2O_3 (%), SiO_2 (%), K_2O (%), CaO (%),
TiO_2 (%), Fe_2 O_3 (%), MnO (ppm), CuO (ppm), ZnO (ppm), PbO_2 (ppm),
Rb_2O (ppm), SrO (ppm), Y_2O_3 (ppm), ZrO_2 (ppm), P_2O_5 (ppm).

Die Scherben stammen aus verschiedenen Epochen der chinesischen Geschichte und von zwei verschiedenen Orten, nämlich Longquan and Jindgezhen. Das Merkmal Ceramic Name kodiert die geographische Herkunft und die Epoche. Relevant für die vorliegende Diskussion ist nur die Herkunft erkennbar am Namensbestandteil FLQ (Longquan) und DY (Jingdgezhen): Diese soll anhand der chemischen Zusammensetzung entweder der Glasur oder des Körpers ermittelt werden. Will man das entstehende Klassifikationsproblem

Abbildung 35: Links: Vase, Ming-Dynastie, Ära Hongwu (1368-1398)
Rechts: Porzellanscherben vom Jingdezhen-Markt
(Quelle links: `http://german.xinhuanet.com`)
(Quelle rechts: Fotografie der Historikerin Anne Gerritsen, Mai 2013)

mit einer Fisher-Klassifikationsregel lösen, so muss also eine affine Funktion mit insgesamt 18 Koeffizienten aus einer Datenmenge mit 44 Samples geschätzt werden. Eine Auswahl geeigneter Merkmale erscheint daher angemessen.

Da die Merkmale stark unterschiedliche Wertebereiche aufweisen, werden alle auf Mittelwert 0 und Standardabweichung 1 transformiert (»standardisiert«).

In Abbildung 36 sind die mit Leaving One Out geschätzten Trefferwahrscheinlichkeiten der 17 möglichen Fisher-Diskriminanzfunktionen mit einem Merkmal bei Betrachtung des Porzellankörpers (Part=Body) dargestellt. Man beachte hierbei, dass die Aufteilung der Datenmenge in Trainings- und Testanteil beim Leaving One Out nach Definition konstant ist. Man erkennt, dass man alleine mit den Merkmalen Fe_2O_3 oder ZrO_2 sehr gute Ergebnisse erhält.

Abbildung 36: Geschätzte Trefferwahrscheinlichkeiten der Fisher-Klassifikation der Datenmenge »Ceramic« bei Nutzung eines Merkmals

Um den möglichen Einfluss der restlichen Merkmale auf die Klassifikation nach dem Herkunftsort zu untersuchen, wurden die beiden dominanten Merkmale Fe_2O_3 und ZrO_2 aus der Betrachtung ausgeschlossen und es wurde in der verbleibenden Menge von 15 Merkmalen eine Best Subset Selection durchgeführt, wobei nur 2- und 3-elementige Teilmengen eingeschlossen wurden. Deren Anzahl beträgt

$$\binom{15}{2} + \binom{15}{3} = 105 + 455 = 560.$$

In Abbildung 37 ist das Ergebnis dargestellt: Man erkennt, dass es eine einzige, optimale Merkmalskombination aus zwei Merkmalen gibt, deren Trefferwahrscheinlichkeit bei etwa 97% liegt. Diese wird auch nicht durch Dreier-Kombinationen von Merkmalen übertroffen. Die Abbildung wurde mit Rapidminer mit Hilfe des Operators »Loop Attributes« erstellt. Dieser liefert allerdings keine Angaben zur jeweiligen Merkmalskombination, sondern nummeriert diese nur. Eine genauere Analyse mit dem Operator »Optimize Selection« ergibt, dass die optimale Zweier-Kombination aus den Merkmalen Rb_2O und Y_2O_3 besteht und eine geschätzte Trefferwahrscheinlichkeit von 97.7% besitzt. Vergleicht man dieses Ergebnis mit der Abbildung 36, so fällt die interessante Tatsache auf, dass die optimale Zweier-Kombination *nicht* aus einer Kombination von Merkmalen besteht, die optimale Diskriminanzfunktionen mit einem Merkmal liefern. Dies wären die Merkmale Al_2O_3, SiO_2, K_3 und CaO. ◇

Forward Selection

Der Rechenaufwand bei Anwendung der Methode Best Subset Selection steigt exponentiell mit der Anzahl p von Merkmalen: Es gibt insgesamt $2^p - 1$ nicht-leere Teilmengen von $\{X_1, \ldots, X_p\}$. In der Praxis müssen also bei hohen Merkmalszahlen, wie sie beispielsweise bei der Analyse von Datenmengen aus dem Bereich der Chemie oder der Genetik auftreten, restriktivere Auswahlverfahren verwendet werden. Eine solche Methode ist die Forward Selection von Merkmalen, bei der man von der gleichen Situation ausgeht wie im Fall der Best Subset Selection.

Abbildung 37: Geschätzte Trefferwahrscheinlichkeiten der Fisher-Klassifikation der Datenmenge »Ceramic« bei Nutzung von zwei oder drei Merkmalen

Die Einzelschritte des Verfahrens sind:

1. Wahl einer maximalen Merkmalsanzahl $1 \leq q_{\max} \leq p$.
2. Optionale Wahl einer zusätzlichen Abbruchbedingung wie zum Beispiel einer Mindestverbesserung der geschätzten Trefferwahrscheinlichkeit durch Hinzunahme eines Merkmals.
3. Initialisierung: Starte mit der leeren Merkmalsteilmenge $T = \emptyset$.
4. Zu jeder Teilmenge $T' \subseteq \{X_1, \ldots, X_p\}$ der Gestalt $T \cup \{X_j\}$, $X_j \notin T$: Bestimmung des Klassifikators $f_{T',\Lambda}$ und seiner geschätzten (gewichteten) Trefferwahrscheinlichkeit $\widehat{P}(f_{\Lambda,T'}(X_1, \ldots, X_p) = Y)_w$.
5. Aktualisierung: Ersetze T durch eine derjenigen Teilmengen T'_0, für die $\widehat{P}(f_{\Lambda,T'_0}(X_1, \ldots, X_p) = Y)_w$ maximal ist.

6. Abbruch: Beende das Verfahren, falls $|T| = q_{\max}$ oder im vorigen Schritt die zusätzliche Abbruchbedingung erfüllt ist, andernfalls führe Schritt 4 erneut durch.

Beispiel 7.2 (Forts. Beispiel 7.1): Es werden erneut die 15 weniger dominanten Merkmale aus Beispiel 7.1 betrachtet. Statt einer Best Subset Selection wird dieses Mal mit der Forward Selection eine optimale Merkmalskombination für die Fisher-Klassifikation gewählt. Die Abbildung 38 zeigt das Ergebnis in einer anderen Darstellung als im Beispiel 7.1. Die Farben kodieren den jeweiligen Wahlschritt in der Forward Selection. So sind zum Beispiel die Ergebnisse der im dritten Schritt des Verfahrens zu testenden $15 - 2 = 13$ Merkmale in der Farbe orange dargestellt.

Die Grafik zeigt, dass das Verfahren insgesamt fünf Merkmale in folgender Reihenfolge und mit den in Klammern angegebenen geschätzten Trefferwahrscheinlichkeiten wählt:

Al_2O_3 (81.8%), Rb_2O (90.9%), K_2O (93.2%), CaO (95.5%), Y_2O_3 (97.7%).

Es ist zu beachten, dass das in Rapidminer implementierte Verfahren im Fall, dass mehrerer Merkmale bei Hinzufügen dieselbe Trefferwahrscheinlichkeit liefern, das jeweils erste wählt.

Nach dem sechsten Schritt endet das Verfahren, weil das Hinzufügen eines sechsten Merkmals keine Erhöhung der Trefferwahrscheinlichkeit mehr erzeugt.

Die Forward Selection kommt im vorliegenden Fall notwendigerweise zu einem anderen Ergebnis mit mehr Merkmalen als die Best Subset Selection, weil im ersten Verfahrensschritt eines der Merkmale (bei Rapidminer das erste) Al_2O_3, SiO_2, K_3 und CaO gewählt wird. Diese Wahl kann im weiteren Verlauf des Verfahrens nicht mehr revidiert werden. Offensichtlich ist diese Eigenschaft ein Nachteil der Forward Selection; er ist der Preis für die kürzere Laufzeit im Vergleich mit der Best Subset Selection.

Abschließend sei noch angemerkt, dass die Datenmenge »Ceramics« eine in Bezug auf die gegebene Problemstellung einfach zu analysierende Datenemnge ist. Dies zeigen zum Beispiel Plots der Ausprägungen in der Stichprobe für jeweils zwei Merkmale, wie in der Abbildung 39 zu sehen. ◇

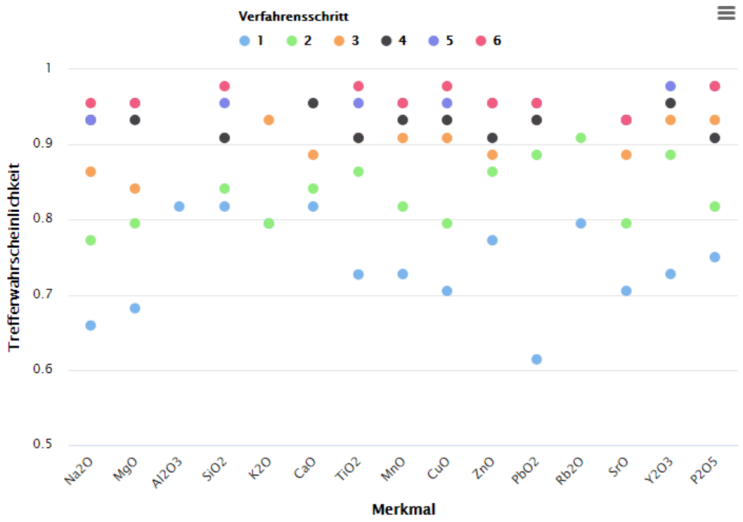


Abbildung 38: Geschätzte Trefferwahrscheinlichkeiten der Fisher-Klassifikation der Datenmenge »Ceramic« im Lauf einer Forward Selection

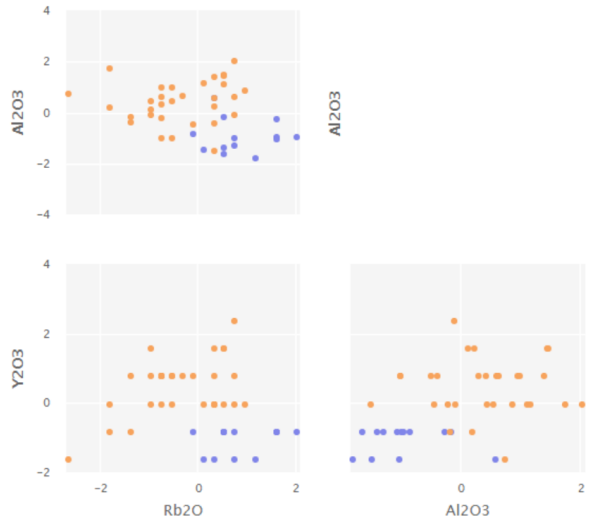


Abbildung 39: Paarweise Scatterplots für die Merkmalsausprägungen von Al_2O_3, Rb_2O und Y_2O_3 im Porzellankörper, sowie der Herkunftsorte (Longquan: blau, Jingdgezhen: orange)

7.2 Merkmalstransformation

Als *Merkmalstransformation* bezeichnet man die Erzeugung eines neuen Merkmals aus einem oder mehreren der Merkmale $X_1, \ldots, X_p$ eines vorliegenden Klassifikationsproblems. Genauer seien $\{X_{k_j} : j \in \{1, \ldots, q\}\}$ eine Teilmenge der vorliegenden Merkmalsmenge, $(T, \mathcal{T})$ ein Messraum und

$$g : S_{k_1} \times \ldots \times S_{k_q} \to T$$

eine messbare Abbildung. Dann ist die Abbildung

$$g(X_{k_1}, \ldots, X_{k_q}) : \Omega \to T,\ \omega \mapsto g(X_{k_1}(\omega), \ldots, X_{k_q}(\omega))$$

messbar, also eine T-wertige Zufallsvariable, die daher als Merkmal der Objekte $\omega \in \Omega$ aufgefasst werden kann.

Besonders häufig tritt der Fall $q = 1$ auf, in dem keine Kopplung mehrerer Merkmale erzeugt wird, sondern nur die Merkmalsausprägungen eines einzelnen Merkmals X_{k_1} transformiert werden.

Die folgenden beiden Merkmalstransformationen sind Standardoperationen beim Umgang mit reellwertigen Merkmalen:

NORMALISIERUNG: Sind alle Merkmale $X_1, \ldots, X_p$ reellwertig mit einem jeweils beschränkten Wertebereich S_k, so kann es aus Vergleichbarkeitsgründen nützlich sein diese so zu transformieren, dass die transformierten Merkmale alle einen Wertebereich mit denselben Schranken besitzen. Genauer seien $a_k := \min(S_k)$ und $b_k := \max(S_k)$; der angestrebte gemeinsame Wertebereich sei das Intervall $[a, b]$. Dann besitzen die transformierten Merkmale

$$X_{k,\mathrm{nor}} : \Omega \to [a, b], \omega \mapsto \frac{b - a}{b_k - a_k}(X_k(\omega) - a_k) + a$$

die gewünschte Eigenschaft.

In der Praxis sind die Werte a_k und b_k entweder aufgrund von Kontextwissen bekannt oder sie müssen anhand einer vorliegenden Stichprobe geschätzt werden. Im letzteren Fall ist zu beachten, dass diese Stichprobe keine Ausreißer enthält, da diese die Schätzung von Minimum und Maximum naturgemäß extrem stören.

STANDARDISIERUNG: Anstatt die reellwertigen Merkmale $X_1, \ldots, X_p$ auf denselben Wertebereich zu transformieren, ist es häufig nützlicher sie so zu

transformieren, dass sie statistisch vergleichbar sind, also denselben Erwartungswert und dieselbe Varianz besitzen. Üblicherweise wählt man dabei den Erwartungswert 0 und die Varianz 1. Dies wird durch die Transformation

$$X_{k,\mathrm{st}} : \Omega \to [a, b], \omega \mapsto \frac{X_k(\omega) - \mathrm{E}(X_k)}{\mathrm{S}(X_k)},$$

erreicht, wobei $\mathrm{E}(X_k)$ den Erwartungswert und $\mathrm{S}(X_k)$ die Standardabweichung bezeichnet. Beide Größen müssen in der Praxis natürlich aus einer vorliegenden Stichprobe geschätzt werden.

7.3 Erweiterung der Merkmalsbasis

In der ursprünglich durch ein Klassifikationsproblem gegebenen Merkmalsmenge Merkmale $X_1, \ldots, X_p$ können einige oder alle Merkmale durch transformierte Merkmale ersetzt werden und man arbeitet zur Lösung des Klassifikationsproblems mit der entstehenden neuen Menge von Merkmalen. Dies ist zum Beispiel bei Normalisierung oder Standadrdisierung der Merkmale der Fall. Es kann jedoch auch nützlich sein durch Transformation entstehende Merkamle zu den ursprünglichen hinzuzunehmen. Man spricht dann von einer *Erweiterung der Merkmalsbasis*. Das folgende Beispiel illustriert den Nutzen dieser Vorgehensweise.

Beispiel 7.3: Man betrachtet die in Abbildung 40 links dargestellte, künstlich erzeugte Datenmenge. Sie besteht aus 2500 Ausprägungen zweier reellwertiger Merkmale X_1 und X_2. Es gibt $r = 2$ Klassen. Aufgrund der Gestalt der beiden Klassen erscheint die Ermittlung eines auf einer affinen Diskriminanzfunktion basierenden Klassifikators nicht intuitiv, obwohl ein mit der Methode von Fisher ermittelter solcher Klassifikator eine geschätzte Trefferwahrscheinlichkeit (10-fache Kreuzvalidierung) von 92% besitzt. Der Klassifikator ist in Abbildung 40 rechts anhand der Darstellung der Klassifikationsergebnisse von 7000 zufällig gewählten Punkten visualisiert.

Die geometrische Anschauung legt nahe, dass eine die beiden Klassen trennende Kurve S-förmig sein sollte, etwa wie der Grapgh einer kubischen Funktion. Aufgrund der räumlichen Lage der Datenmenge im Koordinatensystem liegt es also nahe eine Diskriminanzfunktion der Gestalt

$$d(X_1, X_2) = a_{13}X_1^3 + a_{12}X_1^2 + a_{11}X_1 + a_{21}X_2 + b$$

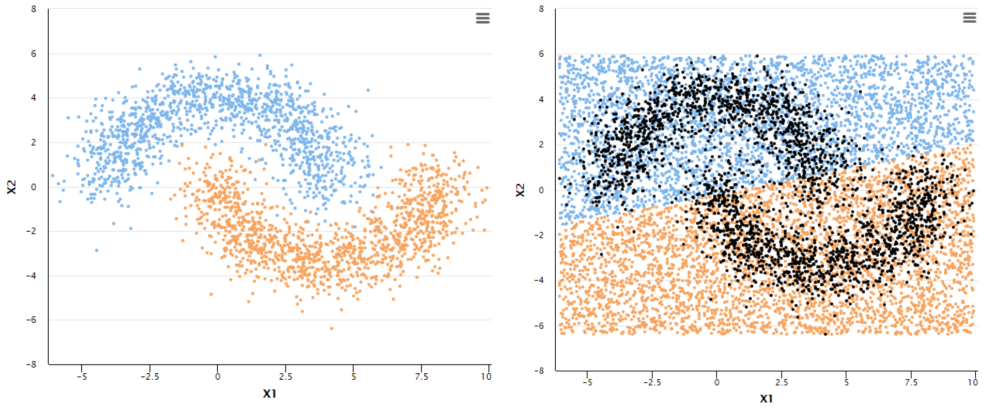


Abbildung 40: Lineare Diskriminanzanalyse einer künstlichen Datenmenge

mit reellen Koeffizienten zu betrachten, die aus den vorliegenden Daten so geschätzt werden müssen, dass die durch die Gleichung $d(X_1, X_2) = 0$ gegebene Kurve die beiden Klassen optimal trennt. Anstatt nun ein Verfahren zur Schätzung dieser Koeffizienten zu entwickeln, kann man diese mit folgender Überlegung auf die Methode von Fisher zurückführen: Man führt die zusätzlichen, durch Transformation aus X_1 entstehenden Merkmale

$$X_3 := X_1^2, \ X_4 := X_1^3$$

ein und führt mit den Merkmalen X_1, X_2, X_3, X_4 eine lineare Diskriminanzanalyse nach Fisher durch. Dabei benutzt man die Datenmenge, deren Samples durch entsprechende Erweiterung der Samples $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$ entstehen, nämlich

$$(x_1^{(i)}, x_2^{(i)}, x_3^{(i)} := (x_1^{(i)})^2, x_4^{(i)} := (x_1^{(i)})^3, y^{(i)}) \in \mathbb{R}^4.$$

Der durch dieses Vorgehen entstehende Klassifikator ist in Abbildung 41 wiederum durch Darstellung der Klassifikationsergebnisse von 7000 zufällig gewählten Punkten visualisiert. Er besitzt eine geschätzte Trefferwahrscheinlichkeit von 97%.

Offensichtlich kann man das im vorangehenden Beispiel beschriebene Vorgehen verallgemeinern:

Quasilineare Diskriminanzanalyse
Es liege die im Abschnitt 5.1 beschriebene Situation im Zwei-Klassen-Fall $(r = 2)$ vor. Insbesondere sind alle Merkmale $X_1, \ldots, X_p$ reellwertig.

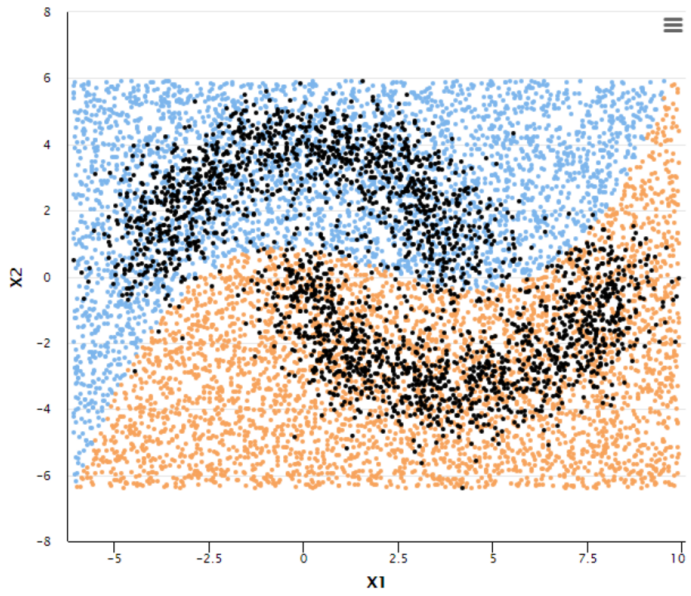


Abbildung 41: Kubische Diskriminanzanalyse einer künstlichen Datenmenge

Es seien $g_k : S_X \to \mathbb{R}$, $k \in \{1, \ldots, q\}$ messbare Funktionen.
Man betrachte den Raum $\mathbf{F}$ von Klassifikatoren der Form

$$f(X_1, \ldots, X_p) = \operatorname{sgn}(b + \sum_{k=1}^{q} a_k g_k(X_1, \ldots, X_p)) \tag{47}$$

mit reellen Koeffizienten $b, a_1, \ldots, a_q$. Dann kann man diese Koeffizienten mit Fisher's Methode anhand einer Stichprobe schätzen, indem man wie folgt vorgeht:

1. Man betrachtet anstelle der ursprünglichen Merkmale $X_1, \ldots, X_p$ die transformierten Merkmale

$$X_1' := g_1(X_1, \ldots, X_p), \ldots, X_q' := g_q(X_1, \ldots, X_p).$$

2. Man transformiert die Samples $(x_1^{(i)}, \ldots, x_p^{(i)}, y^{(i)})$ der vorliegenden Datenmenge zu den Samples

$$(g_1(x_1^{(i)}, \ldots, x_p^{(i)}), \ldots, g_q(x_1^{(i)}, \ldots, x_p^{(i)}), y^{(i)}).$$

3. Man bestimmt mit der Methode von Fisher eine lineare Diskriminanzfunktion

$$d(X'_1, \ldots, X'_q) = b + \sum_{k=1}^{q} a_k X'_k$$

für die transformierten Merkmale und anhand der im Schritt 2 entstandenen Datenmenge.

4. Rückersetzen der Merkmale X'_k durch $g_k(X_1, \ldots, X_p)$ liefert den gesuchten Klassifikator (47).

Polynomiale Diskriminanzanalyse

Ein wichtiger Spezialfall der quasilinearen Diskriminanzanalyse tritt auf, wenn man für die Funktionen g_k Monome in den Merkmalen wählt, also Funktionen der Gestalt

$$g_k(X_1, \ldots, X_p) = X_1^{e_1} \cdot X_2^{e_2} \cdot \ldots \cdot X_p^{e_p},$$

wobei die Exponenten $e_j \in \mathbb{N}_0$ von k abhängen können. Die entstehende Diskriminanzfunktion $d(X_1, \ldots, X_p)$ ist dann ein Polynom in p Variablen, weswegen man von polynomialer Diskriminanzanalyse spricht. Das Beispiel 7.3 ist in diesem Sinn eine polynomiale Diskriminanzanalyse dritten Grades.

Durch Transformation entstandene Merkmale der Gestalt $X_j X_k$ bezeichnet man auch als *Interaktionen*. Das folgende Beispiel macht diese Namensgebung klar:

Beispiel 7.4: Mit einem Kunststoffextruder (Abbildung 42) können zum Beispiel Rohre produziert werden. Im Extruder wird hierzu Kunststoffgranulat über ein Heizsystem erhitzt und gleichzeitig durch eine Förderschnecke in Richtung des Werkzeugs transportiert. Beim Durchpressen des dann plastisch gewordenen Granulats durch das Werkzeug entsteht das Rohr.

Eine Qualitätsgröße für produzierte Rohre ist die Varianz σ^2 der Rohrwanddicke über einen definierten Bereich des Rohres. Sie hängt sowohl von der am Extruder eingestellten Temperatur T, als auch von der Geschwindigkeit v der Förderschnecke ab: Bei zu tiefer Temperatur ist das Granulat, wenn es am Werkzeug ankommt, noch nicht hinreichend stark geschmolzen, wodurch eine ungleichmäßige, sogar mit Granulateinschlüssen versehene Rohrwand entsteht. Ob dieses Verhalten auftritt, hängt jedoch bei gegebener Temperatur auch von der Fördergeschwindigkeit ab, da bei langsamer Schneckengeschwindigkeit das Granulat mehr Zeit zum Schmelzen hat. Zur

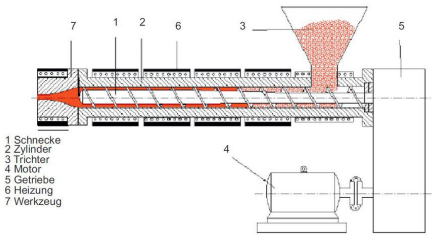


Abbildung 42: Kunststoffextruder

einfachen Modellierung der Größe σ^2 in Abhängigkeit von T und v liegt daher eine Gleichung der Form

$$\begin{aligned} \sigma^2 &= (b_T + a_T T)(b_v + a_v v) \\ &= b_T b_v + b_T a_v v + b_v a_T T + a_T a_v T v \end{aligned}$$

mit reellen Koeffizienten nahe: Bei konstanter Schneckengeschwindigkeit besteht ein linearer Zusammenhang zwischen T und σ^2; bei konstanter Temperatur besteht ein linearer Zusammenhang zwischen v und σ^2.

Ist σ_0^2 die maximal zulässige Varianz für die Produktion von Rohren, so kann man mit dem Klassifikator

$$f(T, v) = \operatorname{sgn}(\sigma_0^2 - b_T b_v - b_T a_v v - b_v a_T T - a_T a_v T v)$$

prognostizieren, ob eine bestimmte Einstellung des Extruders Ausschuss produziert oder verkaufbare Kunststoffrohre.

Die gesamte Darstellung der Problematik ist natürlich stark vereinfacht und daher realitätsfern. ◇

Die polynomiale Diskriminanzanalyse zeigt auch einen möglichen Nachteil der quasilinearen Diskriminanzanalyse auf: Die Anzahl verschiedener Monome in p Variablen und vom Grad kleiner gleich d ist gleich dem Binomialkoeffizienten $\binom{p+d}{p}$. Diese Anzahl wächst mit steigender Merkmalszahl p oder steigendem Grad d schnell an, was damit auch für die Anzahl der in der Diskriminanzfunktion beziehungsweise dem Klassifikator (47) auftretenden Koeffizienten gilt. Der Umfang der nutzbaren Datenmenge bleibt aber gleich, womit die Güte der Schätzung der Koeffizienten immer schlechter wird. Dies legt nahe die Methode der Erweiterung der Merkmalsbasis gegebenenfalls mit einer Methode zur Merkmalsauswahl zu koppeln.

8 Data Mining als Arbeitsprozess

Werden in einem Unternehmen häufiger Aktivitäten oder Projekte im Bereich des Data Mining durchgeführt, so erfolgt dies professionell in mehr oder weniger standardisierten Schritten. Ein solches Vorgehen ist aus verschiedenen Gründen vorteilhaft:

- Verankern von Erfahrungen im standardisierten Prozess (Wissenstradition im Unternehmen),
- Vermeidung von (groben) Verfahrensfehlern,
- gute Dokumentierbarkeit eines Projekts,
- Ermöglichen einer Qualitätssicherung,

um nur einige zu nennen. Die für den Data-Mining-Arbeitsprozess verwendeten Standards variieren natürlich von Unternehmen zu Unternehmen, sind aber in aller Regel an allgemeine Prozessmodelle wie etwa dem Data-Mining-Prozess nach Fayyad, Piatetsky-Shapiro and Smyth oder dem »Cross-Industry Standard Process« (CRISP) angelehnt. Die folgenden Ausführungen beziehen sich auf den erstgenannten Prozess – siehe hierzu den grundlegenden Artikel [F-PS-S] und Abbildung 43. In diesem wird der Prozess des »Knowledge Discovery in Databases« (KDD) beschrieben. Data Mining ist ein Teilschritt dieses Prozesses. Angesichts der Tatsache, dass anspruchsvolle mathematische Methoden auch in anderen Teilschritten des KDD-Prozesses vorkommen, wird inzwischen häufig nicht (mehr) zwischen KDD und Data Mining unterschieden, was auch im Folgenden so gehandhabt wird.

Um einen Arbeitsprozess zu definieren, muss klar sein welches Ziel mit diesem Prozess erreicht werden soll. Eine allgemein anerkannte Beschreibung des Ziels im Falle eines Data-Mining-Prozesses ist folgende: Aus vorliegenden Daten sollen valide, bislang unbekannte, potentiell nützliche und im jeweiligen Anwendungskontext interpretierbare Muster/Strukturen extrahiert werden. Unter »Mustern« oder »Strukturen« versteht man dabei jede Beschreibung der Daten auf einem höheren Niveau als dem der Angabe von Werten und deren Häufigkeiten. Muster sind also zum Beispiel Wahrscheinlichkeitsverteilungen, die aus den Daten geschätzt wurden, oder Regressionsfunktionen durch eine Datenmenge. »Valide« bedeutet, dass die ermittelten Muster mit hinreichend hoher Wahrscheinlichkeit auch in neuen Daten beobachtbar sind, das heißt die Muster spiegeln eine reale Gesetzmäßigkeit wider und sind keine

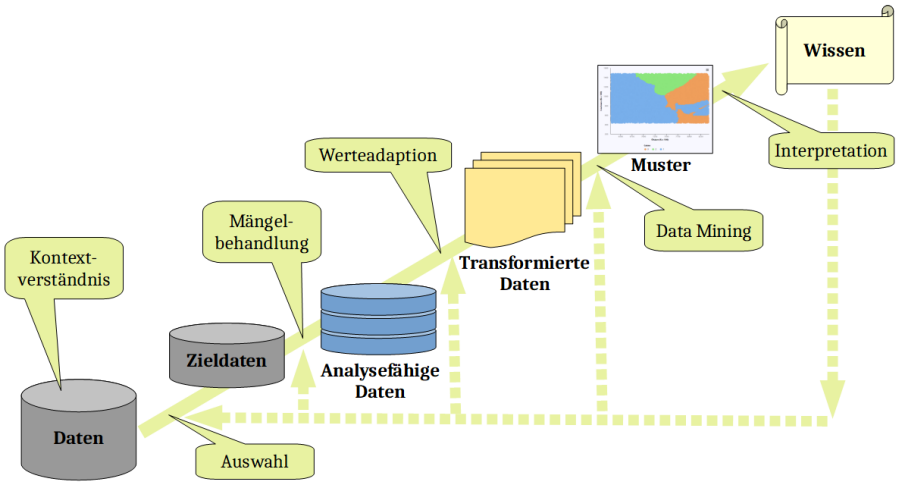


Abbildung 43: Data-Mining-Prozess nach Fayyad, Piatetsky-Shapiro, Smyth

Artefakte. Die ermittelten Muster liefern nach Bewertung und Interpretation durch Menschen möglicherweise Wissen über den Bereich, aus dem die betrachteten Daten stammen.

Nach dem Modell von Fayyad, Piatetsky-Shapiro und Smyth ist der Arbeitsprozess, mit dessen Hilfe man Wissen aus Daten gewinnt, in sechs Teilschritte gegliedert. Diese werden im Folgenden beschrieben. Man beachte dabei, dass der Prozess interaktiv und iterativ ist: Die nach jedem Schritt erhaltenen Ergebnisse müssen bewertet werden; das Bewertungsergebnis kann dazu führen, dass man einen oder mehrere Schritte zurückgehen und diese wiederholen muss.

1. **Verständnis des Kontexts:** Jedes Data-Mining-Projekt findet in einem konkreten Anwendungsrahmen statt und verfolgt anwendungsspezifische Ziele. Letztere müssen zusammen mit dem nötigen Kontextwissen von den Durchführenden des Projekts verstanden werden.

   Beispiel: Daten einer Anlage zur Produktion von Kunststoffspritzgussteilen sollen analysiert werden. Allgemeines Ziel ist die Verbesserung der Qualität der produzierten Teile. Anwendungsspezifisches Ziel: Es kommt besonders auf die farbliche Homogenität der Teile an, da sie im Innendesign von Wohnräumen zum Einsatz kommen. Kontextwissen: Wie und mit welcher Genauigkeit wird die Farbhomogenität ermittelt?

Durch welche Einstellungen an der Spritzgussanlage kann sie beeinflusst werden? Was kosten geänderte Einstellungen?

2. **Datenauswahl:** Es wird entschieden welche Daten und welche Merkmale dieser Daten in der Analyse verwendet werden. Je nach Kontext liegen die Daten bereits vor oder werden kontinuierlich ermittelt. Falls dies nicht der Fall ist, wird eine Methode zur Erhebung dieser Daten festgelegt und die Datenerhebung wird durchgeführt. In diesem Arbeitsschritt werden oft statistische Überlegungen und Tests genutzt, um notwendige Entscheidungen zu unterstützen.

   Beispiel: Es wird entschieden alle über die Zeit protollierten Einstellungen der Spritzgussanlage zu verwenden, da man nicht genau weiß, welche Einstellungen die Farbhomogenität beeinflussen. Letztere wird bislang vom Maschinenführer optisch regelmäßig geprüft und handschriftlich protokolliert. Es wird entschieden eine Kamera zu installieren, die in einem bestimmten Zeittakt Nahaufnahmen der produzierten Teile aufnimmt. Diese werden in die Analyse eingeschlossen.

3. **Datenvorverarbeitung:** Die zu analysierenden Daten werden allgemein gesprochen in eine analysefähige Form gebracht. Hierzu sind kontextabhängig zum Beispiel folgende Operationen notwendig:

   - Entrauschen der Daten, das heißt Entfernen oder wenigstens Reduzieren von Artefakten und Störungen in den Daten, die durch den Erhebungsvorgang verursacht werden.
   - Festlegen einer Methode mit fehlenden Daten umzugehen, etwa bei temporärem Ausfall der Datenerfassung oder bei Fehlen einzelner Merkmalsausprägungen in Datensätzen.
   - Festlegen einer Methode der Zusammenführung von Daten aus verschiedenen Datenquellen, etwa bei zeitabhängigen Daten, die mit unterschiedlicher Frequenz ermittelt werden.

   Beispiel: Die Temperatur der Kunststoffmasse wird nach dem Füllen der Form an verschiedenen Stellen im Spritzgusswerkzeug mit Sensoren ermittelt. Es kommt vor, dass die Form nicht vollständig gefüllt wird, die Temperaturwerte an entsprechenden Stellen als nicht valide sind. Die in einem solchen Fall produzierten Teile sind Ausschuss, obwohl ihre Farbhomogenität gut sein kann. Prinzipiell sind die Daten zu solchen

Ausschussteilen also interessant. Es wird aber entschieden wegen des teilweisen Fehlens valider Temperaturwerte jeden Datensatz zu einem Ausschussteil nicht in die Analyse aufzunehmen.

4. **Datentransformation:** Es ist häufig zweckmäßig oder notwendig die Datensätze vor der Analyse in für das angestrebte Ziel geeignetere Datensätze zu transformieren. Häufige Ziele der Datentransformation sind:

   - Erzeugen vergleichbarer Wertebereiche bei numerischen Merkmalen durch Normalisieren oder Standardisieren.
   - Umwandeln eines Merkmals vom vorliegenden Typ in einen anderen: Die Diskretisierung eines numerischen Merkmals in ein nominales ist ein Beispiel dieser Operation.
   - Vereinfachen von komplexen Merkmalen: Besitzt ein Merkmal beispielsweise Zeitreihen als Ausprägungen, so verwendet man von diesen nach der Transformation nur ihren Mittelwert und die Varianz.
   - Reduktion der Anzahl zu betrachtender Merkmale durch Kombination mehrerer Merkmale zu einem neuen (so genannte »Dimensionsreduktion«).

   Beispiel: Der während eines einzelnen Spritzgussvorgangs (Produktion eines Teils) gemessene Druckverlauf im Werkzeug wird durch den mittleren Druck während des Gussvorgangs ersetzt. Das regelmäßig von der Kamera aufgezeichnete Bild wird durch die Anzahl der signifikanten Farbinhomogenitäten, sowie durch den Wert der größten vorkommenden Farbinhomogenität ersetzt (beachte: Bilder sind eine Matrix von Farbwerten z.B. in der Menge $\{0, \ldots, 255\}$ (RGB-Kode)). Im letzten Fall werden aus einem komplexen Merkmal zwei numerische erzeugt.

5. **Data Mining:** Die nun vorliegenden analysefähigen und geeignet transformierten Daten werden in diesem Schritt analysiert. Die Analyse zerfällt in mehrere Teilschritte, die alle ein erhebliches Fachwissen in den Bereichen Statistik, Data Mining und automatisches Lernen erfordern:

   - Wahl einer passenden Data-Mining-Methode,
   - Wahl eines geeigneten Algorithmus', der die Methode realisiert,

- Bestimmung eines Modells der Daten, das heißt Suche nach und Beschreibung von Mustern/Strukturen in den Daten,
- Validierung des Modells.

Beispiel: Mit Hilfe von nichtlinearer Regression (Data-Mining-Methode) wird unter Verwendung eines neuronalen Netzes (Algorithmus) ein näherungsweiser Zusammenhang zwischen der Kunststofftemperatur und dem mittleren Druckverlauf während des Gussvorgangs einerseits und der Farbhomogenität andererseits ermittelt.

6. **Bewertung und Interpretation:** Die aus den Daten ermittelten Modelle werden im Anwendungskontext interpretiert und es wird untersucht, inwieweit sie etwas zu den spezifischen Zielen des Projekts beitragen. Bei Data-Mining-Projekten besteht generell das Risiko, dass die Ergebnisse im Anwendungskontext irrelevant oder nicht umsetzbar sind. Dieses Risiko führt oft dazu, dass kleine und mittelständische Unternehmen Data-Mining-Projekte wegen des hohen Kostenrisikos erst gar nicht ins Auge fassen.

   Beispiel: Das Regressionsmodell für die Farbhomogenität kann genutzt werden, um Einstellungen der Spritzgussanlage zu ermitteln, bei denen die Farbhomogenität (näherungsweise) optimal ist. Es ist dann zu diskutieren, ob diese Einstellungen andere Qualitätsgrößen negativ beeinflussen oder ob die Betriebskosten der Anlage zum Beispiel durch erhöhten Energiebedarf steigen. Es ist zu entscheiden, ob Negativeffekte sich in einem akzeptablen Rahmen halten oder nicht.

Literatur

[Bre] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees. Wadsworth, New York 1984.

[Dar] C. Darwin, Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, 1845.

[F-PS-S] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine 17 (3) (1996).

[FAL] M. Forina, C. Armanino, S. Lanteri, Classification of olive oils from their fatty acid composition, Food Research and Data Analysis, 1983.

[Fis] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936), 179–188.

[FHT] J. Friedman, T. Hastie, R. Tibshirani, Introduction to Statistical Learning Theory, Springer 2008.

[WFH] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann 2017.

[Geo] H.-O. Georgii, Stochastik, Walter de Gruyter 2004.

[Gie] T. Giersch, Olivenölherstellung – wie im Kokainhandel, Handelsblatt, 7. Oktober 2012.

[Kag] Kaggle, Online-Community für Datenwissenschaftler, `https://www.kaggle.com`

[Spe] Lexikon der Neurowissenschaften, `https://www.spektrum.de`

[RKI] Diabetes Surveillance basierend auf den RKI-Befragungs- und Untersuchungssurveys 1997-1999

(BGS98) und 2008-2011 (DEGS1), Robert-Koch-Institut, https://diabsurv.rki.de

[TK] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Academic Press 2006.

[UCI] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/

[Vap] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York 2000.