MUSTERERKENNUNG

Vorlesung im Sommersemester 2017

Prof. E.G. Schukat-Talamazzini

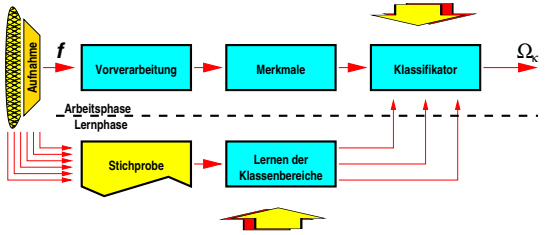
Stand: 30. Mai 2017

Teil VI

Numerische Klassifikation

Numerische Klassifikationsaufgabe

Raten der wahren Musterklasse Ω_κ von $\boldsymbol{f} \in \boldsymbol{\Omega}$



Zielsetzung

Die **geratene** Klasse $\hat{\kappa}$ von $\boldsymbol{x}$ bzw. $\boldsymbol{f}$ ist (möglichst oft) gleich der **wahren** Klasse κ.

Musterrepräsentation

ein Vektor mit numerischen Merkmalen

$$\boldsymbol{x} = (x_1, \ldots, x_D)^\top \in \mathbb{R}^D$$

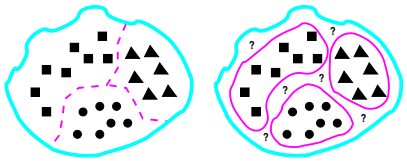
Klassifikationsergebnis

die vermutete Klassenzugehörigkeit

$$\hat{\kappa}(\boldsymbol{x}) \in \{1, 2, \ldots, K\}$$

Entscheidungsgrundlage eines Klassifikators

Modelle für eine harte oder eine weiche Klasseneinteilung



Entscheidungsfunktion

$$\delta : \begin{cases} \mathbb{R}^D & \to & \{1,\ldots,K\} \\ \boldsymbol{x} & \mapsto & \hat{\kappa} = \delta(\boldsymbol{x}) \end{cases}$$

Probabilistisches Modell

Der Klassifikator kennt **Zugehörigkeitsgrade** $\hat{\mathrm{P}}(\kappa|\boldsymbol{x})$ der Muster zu den Klassen Ω_κ.

Geometrisches Modell

Der Klassifikator kennt nur eine **Klassenpartition** $\{\hat{\boldsymbol{\Omega}}_1,\ldots,\hat{\boldsymbol{\Omega}}_K\}$ des Merkmalraums oder ihre Grenzen.

Maschinelles Lernen eines Klassifikators

Lehrer $\hat{=}$ endliche Menge von Beispielmustern

Induktion

Der Klassifikator ist aus einer **endlichen Probe** $\omega \subset \boldsymbol{\Omega}$ des Merkmalraums zu „lernen“.

Lernmethode

- alle Muster aus ω etikettiert — überwacht
- kein Muster aus ω etikettiert — unüberwacht
- einige Muster aus ω etikettiert — teilüberwacht

Das Etikettieren von Mustern ist i.a. *viel teurer* als das Beschaffen selbst!

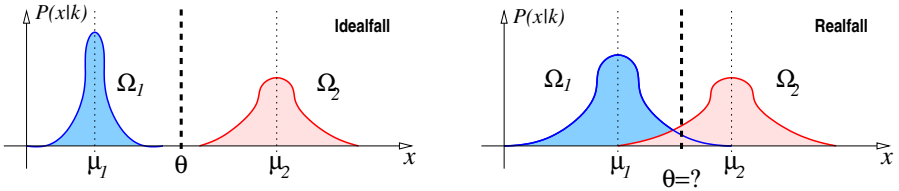
Klassifikatorleistung

Erwartete und beobachtete Fehlerhäufigkeit

Klassifikatortest

Die Fehlerrate einer Entscheidungsregel δ wird mit Hilfe einer (etikettierten) **Teststichprobe** ω' geschätzt:

$$\hat{p}_\varepsilon(\{\omega'_\kappa\}|\delta) \stackrel{\text{def}}{=} \frac{\text{falsch klassifiziert}}{\text{alle Muster}} = \frac{\sum_{\kappa=1}^{K} \#\{\boldsymbol{x} \in \omega'_\kappa \mid \delta(\boldsymbol{x}) \neq \kappa\}}{\sum_{\kappa=1}^{K} |\omega'_\kappa|}$$



Bemerkung

Überlagern sich die Klassengebiete im gewählten Merkmalraum, so können nicht *alle* Muster im Überlappungsbereich korrekt klassifiziert werden!

Statistische Entscheidungstheorie

Viel Wind um einen Papiertiger?

Voraussetzungen

- Die Wahrscheinlichkeitsverteilung des **mustererzeugenden Prozesses** ist bekannt.
- Ein **Gütekriterium** (Kostenfunktion) für mögliche Klassifikationsentscheidungen liegt vor.

Resultat

- Die **Bayesregel** ist der optimale (kostenminimale) Klassifikator.
- Die Bayesregel ist sogar optimal unter allen unscharfen Entscheidungsregeln.

Problem

Die Bayesregel ist leider nicht effektiv berechenbar ...

Generatives Modell

Wahrscheinlichkeitsverteilung für die Mustererzeugung („$1 + K$ Urnen“)

Gemeinsame Verteilung

Ein Muster ist durch seine Klasse $\kappa \in \{1..K\}$ und den Vektor $\boldsymbol{x} \in \mathbb{R}^D$ seiner Merkmale charakterisiert:

$$\mathrm{P}(\kappa) \cdot \mathrm{P}(\boldsymbol{x}|\kappa) \;=\; \mathrm{P}(\boldsymbol{x}, \kappa) \;=\; \mathrm{P}(\boldsymbol{x}) \cdot \mathrm{P}(\kappa|\boldsymbol{x})$$

Randverteilungen

A priori Klassenwahrscheinlichkeiten

$$\mathrm{P}(\kappa) \;=\; \int \mathrm{P}(\boldsymbol{x}, \kappa)\, d\boldsymbol{x}$$

Marginale Merkmalvektordichte

$$\mathrm{P}(\boldsymbol{x}) \;=\; \sum_{\kappa=1}^{K} \mathrm{P}(\boldsymbol{x}, \kappa)$$

Bedingte Verteilungen

Klassenbedingte Merkmalvektordichte

$$\mathrm{P}(\boldsymbol{x}|\kappa) \;=\; \mathrm{P}(\boldsymbol{x}, \kappa) \,/\, \mathrm{P}(\kappa)$$

A posteriori Klassenwahrscheinlichkeiten

$$\mathrm{P}(\kappa|\boldsymbol{x}) \;=\; \mathrm{P}(\boldsymbol{x}, \kappa) \,/\, \mathrm{P}(\boldsymbol{x})$$

Qualitätskriterium

Kostenmodell für die (Fehl-)klassifikation eines Musters

Kostenmatrix

$$\begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,K} \\ \vdots & \cdots & \ddots & \vdots \\ r_{K,1} & r_{K,2} & \cdots & r_{K,K} \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} r_{1,0} & r_{1,1} & r_{1,2} & \cdots & r_{1,K} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ r_{K,0} & r_{K,1} & r_{K,2} & \cdots & r_{K,K} \end{pmatrix}$$

Eine Entscheidung für die Klasse Ω_λ bei Vorlage eines Musters aus Ω_κ verursacht die **Kosten** $r_{\kappa\lambda}$.

- Der Eintrag $r_{\kappa,0}$ beziffert die **Rückweisungskosten**.
- Plausible Forderung ($\lambda \neq \kappa$): $\quad 0 \leq r_{\kappa\kappa} < r_{\kappa 0} < r_{\kappa\lambda}$
- 0/1-Kostenmatrix zur Zählung der Klassifikationsfehler:

$$r_{\kappa\lambda} = \begin{cases} 0 & \lambda = \kappa \\ 1 & \lambda \neq \kappa \end{cases}$$

Entscheidungsregel

Scharfe und randomisierte Entscheidungsregeln

Definition

Für einen Problemkreis mit K Musterklassen und dem Merkmalraum $\mathbb{R}^D$ heißt eine Abbildung

$$\boldsymbol{\delta} : \begin{cases} \mathbb{R}^D & \to & \mathbb{R}^{K+1} \\ \boldsymbol{x} & \mapsto & \boldsymbol{\delta}(\boldsymbol{x}) = (\delta_0(\boldsymbol{x}), \delta_1(\boldsymbol{x}), \ldots, \delta_K(\boldsymbol{x}))^\top \end{cases}$$

Entscheidungsregel, falls die Normierungsbedingung

$$\sum_{\lambda=0}^{K} \delta_\lambda(\boldsymbol{x}) \;=\; 1\,, \quad (\kappa = 1, \ldots, K)$$

gilt. Die Regel heißt **scharf** (hart), wenn alle $\boldsymbol{\delta}(\boldsymbol{x})$ Einheitsvektoren sind; andernfalls heißt die Regel **unscharf** (weich, randomisiert).

Eine unscharfe Regel liefert zum Beispiel Aussagen der Form $\boldsymbol{\delta}(\boldsymbol{x}) = (1/10, 2/10, 7/10, 0)^\top$ oder $\boldsymbol{\delta}(\boldsymbol{x}) = (1/2, 1/8, 1/8, 1/4)^\top$

Risikoformel

Zu erwartende Klassifikationskosten einer Entscheidungsregel

Lemma

*Die zu erwartenden Kosten (das **Risiko**) der Klassifikation von Mustern durch die Entscheidungsregel*

$$\boldsymbol{\delta}(\boldsymbol{x}) : \mathbb{R}^D \;\to\; \mathbb{R}^{K+1}$$

auf Grundlage der Kostenmatrix $[r_{\kappa\lambda}]$ *betragen*

$$\mathfrak{R}(\boldsymbol{\delta}) \;=\; \sum_{\kappa=1}^{K} p_\kappa \sum_{\lambda=0}^{K} r_{\kappa\lambda} \int_{\mathbb{R}^D} \mathrm{P}(\boldsymbol{x}|\Omega_\kappa) \cdot \delta_\lambda(\boldsymbol{x})\, d\boldsymbol{x}$$

Im Falle der 0/1-Kostenmatrix lautet die Risikoformel

$$\mathfrak{R}(\boldsymbol{\delta}) \;=\; \sum_{\kappa=1}^{K} p_\kappa \int_{\mathbb{R}^D} \mathrm{P}(\boldsymbol{x}|\Omega_\kappa) \cdot \sum_{\lambda\neq\kappa} \delta_\lambda(\boldsymbol{x})\, d\boldsymbol{x} \;=\; \sum_{\kappa=1}^{K} p_\kappa \int_{\mathbb{R}^D} \mathrm{P}(\boldsymbol{x}|\Omega_\kappa) \cdot (1 - \delta_\kappa(\boldsymbol{x}))\; d\boldsymbol{x}$$

und quantifiziert die **Fehlerwahrscheinlichkeit** der Entscheidungsregel $\boldsymbol{\delta}(\cdot)$.

Optimale Entscheidungsregel

Allgemeine Kostenmatrix · Rückweisungsmöglichkeit

Satz

Die optimale (randomisierte) Entscheidungsregel $\boldsymbol{\delta}^\star$*, welche das Risiko* $\mathfrak{R}(\boldsymbol{\delta})$ *hinsichtlich einer gegebenen Kostenmatrix* $[r_{\kappa\lambda}]$ *minimiert, klassifiziert die Merkmalvektoren gemäß*

$$\delta^\star_\kappa(\boldsymbol{x}) \;=\; \begin{cases} 1 & u_\kappa(\boldsymbol{x}) = \min_\lambda u_\lambda(\boldsymbol{x}) \\ 0 & \textit{sonst} \end{cases} \quad , \quad \boldsymbol{x} \in \mathbb{R}^D$$

*mit der **Prüfgröße***

$$u_\lambda(\boldsymbol{x}) \;\stackrel{def}{=}\; \sum_{\kappa=1}^{K} r_{\kappa\lambda} \cdot p_\kappa \cdot \mathrm{P}(\boldsymbol{x}|\Omega_\kappa) \;, \quad \lambda = 0, 1, \ldots, K \;.$$

Bemerkung

Diese optimale Entscheidungsregel ist *scharf*!

Beweis.

Das Risiko $\mathfrak{R}(\boldsymbol{\delta})$ setzt sich zusammen als gewichtete Summe der klassenweisen Einzelrisiken:

$$\mathfrak{R}(\boldsymbol{\delta}) \;=\; \sum_{\kappa=1}^{K} p_\kappa \cdot \mathfrak{R}_\kappa(\boldsymbol{\delta})$$

$\mathfrak{R}_\kappa(\boldsymbol{\delta})$ wiederum setzt sich aus den einzelnen Wagnisarten zusammen, die sich aus der jeweiligen Klassifikationsentscheidung ergeben:

$$\mathfrak{R}_\kappa(\boldsymbol{\delta}) \;=\; \sum_{\lambda=0}^{K} r_{\kappa\lambda} \cdot \mathrm{P}(\text{klassifiziert nach } \Omega_\lambda \mid \text{stammt aus } \Omega_\kappa)$$

Den bedingten Wahrscheinlichkeitsausdruck gewinnen wir durch Marginalisierung:

$$\mathrm{P}(\Omega_\lambda|\Omega_\kappa) \;=\; \int_{\mathbb{R}^D} \mathrm{P}(\Omega_\lambda, \boldsymbol{x} \mid \Omega_\kappa)\, d\boldsymbol{x}$$

Mehrmalige Anwendung der Definition bedingter Wahrscheinlichkeiten nebst Erweiterung ergibt:

$$\mathrm{P}(\Omega_\lambda, \boldsymbol{x} \mid \Omega_\kappa) \;=\; \frac{\mathrm{P}(\Omega_\lambda, \boldsymbol{x}, \Omega_\kappa)}{\mathrm{P}(\Omega_\kappa)} \cdot \frac{\mathrm{P}(\boldsymbol{x}, \Omega_\kappa)}{\mathrm{P}(\boldsymbol{x}, \Omega_\kappa)} \;=\; \mathrm{P}(\Omega_\lambda \mid \boldsymbol{x}, \Omega_\kappa) \cdot \mathrm{P}(\boldsymbol{x} \mid \Omega_\kappa) \;=\; \delta_\lambda(\boldsymbol{x}) \cdot \mathrm{P}(\boldsymbol{x} \mid \Omega_\kappa)$$

□

Beweis.

Mit obiger Prüfgrößendefinition gilt die Darstellung

$$\mathfrak{R}(\boldsymbol{\delta}) \;=\; \int_{\mathbb{R}^D} \underbrace{\sum_{\lambda=0}^{K} \underbrace{\left[\sum_{\kappa=1}^{K} r_{\kappa\lambda}\, p_\kappa\, \mathrm{P}(\boldsymbol{x}|\Omega_\kappa)\right]}_{u_\lambda(\boldsymbol{x})} \delta_\lambda(\boldsymbol{x})}_{I(\boldsymbol{x})}\; d\boldsymbol{x}$$

Der Integrand $I(\boldsymbol{x})$ gehorcht der Abschätzung

$$\begin{aligned} I(\boldsymbol{x}) \;&=\; \sum_{\lambda=0}^{K} u_\lambda(\boldsymbol{x}) \cdot \delta_\lambda(\boldsymbol{x}) \\ &\geq\; \sum_{\lambda=0}^{K} u_{\min}(\boldsymbol{x}) \cdot \delta_\lambda(\boldsymbol{x}) \\ &=\; u_{\min}(\boldsymbol{x}) \sum_{\lambda=0}^{K} \delta_\lambda(\boldsymbol{x}) \;=\; u_{\min}(\boldsymbol{x}) \end{aligned}$$

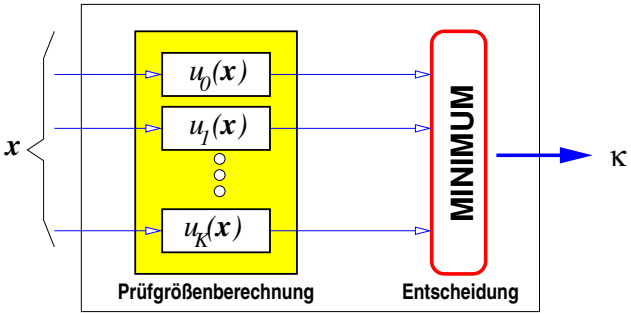
Dieser Minimalwert $u_{\min}(\boldsymbol{x}) = \min u_\lambda(\boldsymbol{x})$ läßt sich durch die Wahl

$$\boldsymbol{\delta}^\star(\boldsymbol{x}) \;=\; \kappa^\star\text{-ter Einheitsvektor}$$

mit $u_{\min}(\boldsymbol{x}) = u_{\kappa^\star}(\boldsymbol{x})$ systematisch erzwingen. □

Der optimale Klassifikator

Allgemeine Kosten · Vereinfachte Kosten



Prüfgröße

$$\underbrace{\sum_{\kappa=1}^{K} r_{\kappa\lambda}\, p_\kappa \mathrm{P}(\boldsymbol{x}|\Omega_\kappa)}_{u_\lambda(\boldsymbol{x})}$$

Vereinfachte Kostenmatrix

$$r_{\kappa\lambda} = \begin{cases} r_w & \kappa = \lambda \\ r_z & \lambda = 0 \\ r_f & \text{sonst} \end{cases} \Rightarrow \quad \begin{aligned} u_0(\boldsymbol{x}) &= r_z \cdot \sum_{\kappa=1}^{K} p_\kappa \cdot \mathrm{P}(\boldsymbol{x}|\Omega_\kappa) = r_z \cdot \mathrm{P}(\boldsymbol{x}) \\ u_\lambda(\boldsymbol{x}) &= r_f \cdot \mathrm{P}(\boldsymbol{x}) + (r_w - r_f) \cdot p_\lambda \cdot \mathrm{P}(\boldsymbol{x}|\Omega_\lambda) \end{aligned}$$

Beweis.

Die vereinfachte Kostenmatrix mit den drei Parametern r_w, r_f, r_z wird bei Klassifikatoren mit erzwungener Entscheidung o.B.d.A. zu

$$r_{\kappa\lambda} = \begin{cases} 0 & \kappa = \lambda \\ 1 & \kappa \neq \lambda \end{cases} .$$

Das Risiko $\Re(\delta)$ entspricht dann der *Fehlerwahrscheinlichkeit* des Klassifikators und seine Prüfgrößen lauten

$$u_\lambda(\boldsymbol{x}) = \sum_{\kappa \neq \lambda} p_\kappa \cdot \mathrm{P}(\boldsymbol{x}|\Omega_\kappa) = \sum_{\kappa \neq \lambda} \mathrm{P}(\boldsymbol{x}, \Omega_\kappa) = \mathrm{P}(\boldsymbol{x}) - \mathrm{P}(\boldsymbol{x}, \Omega_\lambda)$$

Statt diese Prüfgröße zu *minimieren*, können wir auch die gemeinsame Wahrscheinlichkeit $\mathrm{P}(\boldsymbol{x}, \Omega_\lambda)$ von Muster und Klasse *maximieren*, denn der Randdichtewert $\mathrm{P}(\boldsymbol{x})$ ist konstant bezüglich der Wahl von Ω_λ. □

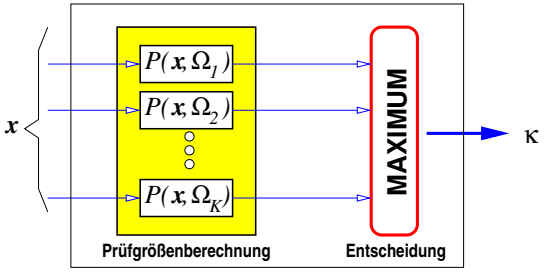
Bemerkung

Herzlichen Dank für dieses interessante *theoretische* Resultat. Aber wie sieht es denn nun eigentlich aus mit der *praktischen* Berechenbarkeit von $\hat{\kappa}(\boldsymbol{x})$ nach der Bayesregel?

1. Die klassenbedingten Musterwahrscheinlichkeiten $\mathrm{P}(\boldsymbol{x}|\Omega_\kappa)$ sind nicht bekannt und nur sehr unzureichend und/oder aufwändig zu modellieren.
2. Die a priori Klassenwahrscheinlichkeiten $\mathrm{P}(\Omega_\kappa)$ sind ebenfalls unbekannt, jedoch sind i.a. leicht robuste Schätzwerte zu beschaffen.
3. Der Randverteilungsdichtewert $\mathrm{P}(\boldsymbol{x})$ ist ebenfalls unbekannt, aber bei Vorliegen der obengenannten Größen unproblematisch durch Aufsummieren der $\mathrm{P}(\boldsymbol{x}, \Omega_\lambda)$ zu berechnen.
4. Zur Klassifikation selbst wird der Wert $\mathrm{P}(\boldsymbol{x})$ allerdings garnicht benötigt, denn er ist ja unabhängig vom Klassenindex κ.

Bayesregel

Klassifikator mit der minimalen Fehlerrate



Satz

Die Bayesregel (Maximum a posteriori-Klassifikator)

$$\delta(\boldsymbol{x}) = \underset{\lambda}{\mathit{argmax}}\, \mathrm{P}(\Omega_\lambda|\boldsymbol{x}) = \underset{\lambda}{\mathit{argmax}}\, \frac{p_\lambda \cdot \mathrm{P}(\boldsymbol{x}|\Omega_\lambda)}{\mathrm{P}(\boldsymbol{x})}$$

ist derjenige Klassifikator mit der kleinstmöglichen Fehlerrate p_ε^{BA} *(**Bayesfehlerrate**).*

Bayesregel

Garantiert minimale Fehlerrate — praktisch nicht verwertbar

Klassifiziere ein neues Muster $\boldsymbol{x} \in \mathbb{R}^D$ gemäß

$$\hat{\kappa}(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\lambda=1..K} u_\lambda(\boldsymbol{x})$$

mit der Prüfgröße

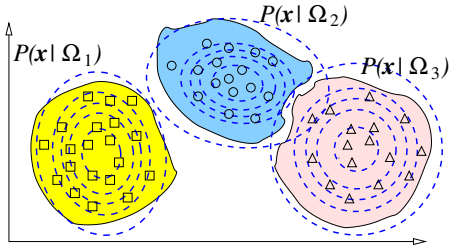
$$u_\lambda(\boldsymbol{x}) \;=\; \mathrm{P}(\Omega_\lambda|\boldsymbol{x}) \;=\; \frac{\mathrm{P}(\Omega_\lambda)\cdot\mathrm{P}(\boldsymbol{x}|\Omega_\lambda)}{\mathrm{P}(\boldsymbol{x})}$$

Typen realer Klassifikatoren

1. Schätze die K Verteilungen $\mathrm{P}(\cdot|\Omega_\lambda)$ — statistisch
2. Schätze die K Trennfunktionen $u_\lambda(\cdot)$ — diskriminativ
3. Rate „auf Zuruf“ die Zugehörigkeit $\hat{\kappa}(\cdot)$ — partitionierend

Statistischer Klassifikator

Schätze Wahrscheinlichkeitsverteilung $f_\kappa(\boldsymbol{x})$ für jede Musterklasse



Modellinformation

- beste Klasse je Muster
- Zugehörigkeitsmaß je Muster
- Verteilungsmodell je Klasse
- explizite Klassengrenzen

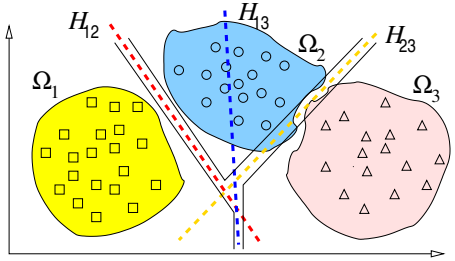
Vorgehensweise

Schätze die wahren klassenbedingten Verteilungsdichten $\mathrm{P}(\boldsymbol{x}|\Omega_\kappa)$ mit Hilfe eines (parametrischen?) Verteilungsmodells $f_\kappa(\boldsymbol{x}) = f(\boldsymbol{x}|\theta_\kappa)$ und entscheide nach der näherungsweisen Bayesregel

$$\delta(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\kappa=1..K} \hat{\mathrm{P}}(\Omega_\kappa|\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\kappa=1..K} \frac{p_\kappa \cdot f_\kappa(\boldsymbol{x})}{\sum_{\lambda=1}^{K} p_\lambda \cdot f_\lambda(\boldsymbol{x})}$$

Diskriminativer Klassifikator

Schätze Klassenzugehörigkeiten oder -trennfunktionen für jedes Muster



Modellinformation

- beste Klasse je Muster
- Zugehörigkeitsmaß je Muster
- explizite Klassengrenzen
- Verteilungsmodell je Klasse

Vorgehensweise

Approximiere geeignete Trennfunktionen

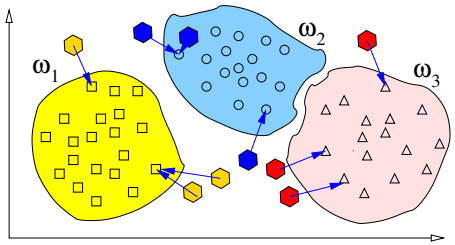
$$h_{\kappa\lambda}(\boldsymbol{x}) \;=\; r \text{ mit } \begin{cases} r \geq 0 & \boldsymbol{x} \in \omega_\kappa \\ r < 0 & \boldsymbol{x} \in \omega_\lambda \end{cases}$$

zur Trennung der Klassengebiete Ω_κ, Ω_λ und setze

$$\delta(\boldsymbol{x}) \;=\; \kappa^\star\,, \quad \text{falls } h_{\kappa^\star\lambda}(\boldsymbol{x}) \geq 0 \text{ für alle } \lambda \neq \kappa^\star$$

Partitionierender Klassifikator

Rate bestpassenden Klassenindex $\hat{\kappa}(\boldsymbol{x})$ für jedes Muster



Modellinformation

- beste Klasse je Muster
- Zugehörigkeitsmaß je Muster
- explizite Klassengrenzen
- Verteilungsmodell je Klasse

Vorgehensweise

Ermittle zum Eingabemuster $\boldsymbol{x}$ den *nächsten Nachbarn*

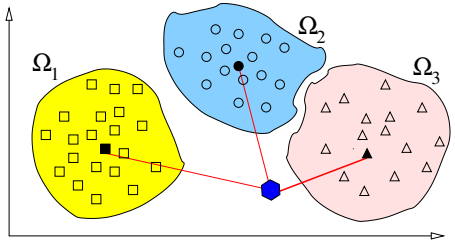
$$\boldsymbol{x}_{\mathrm{NN}} \;=\; \operatorname*{argmin}_{\boldsymbol{z}\in\bigcup\omega_\kappa} \|\boldsymbol{z}-\boldsymbol{x}\|$$

in der Lernstichprobe und setze

$$\delta(\boldsymbol{x}) \;=\; \kappa \quad \text{für} \quad \boldsymbol{x}_{\mathrm{NN}} \in \omega_\kappa$$

Abstandsmessender Klassifikator

Entscheide auf Grundlage der Ähnlichkeiten zwischen Muster und allen Klassenprototypen



Modellinformation

- beste Klasse je Muster
- Zugehörigkeitsmaß je Muster
- Klassengrenzen (Voronoizellen)
- Verteilungsmodell je Klasse

Vorgehensweise

Wähle geeignete Klassenprototypen $\boldsymbol{\mu}_\kappa$ und setze

$$\delta(\boldsymbol{x}) \;=\; \operatorname*{argmin}_{\kappa} \|\boldsymbol{x} - \boldsymbol{\mu}_\kappa\|_q$$

Definition der Prototypen?
Definition des Abstandsmaßes?

Informationshierarchie

Stärkere und schwächere Klassifikatormodelle

Verteilungsfunktionen

Wie sind die Muster einer Klasse über den Raum hinweg gestreut?

???

nicht aus den $u_\kappa(\boldsymbol{x})$ reproduzierbar, da $\mathrm{P}(\boldsymbol{x})$ fehlt

Zugehörigkeiten

Wie verteilt sich die Paßfähigkeit eines Musters auf die Klassen?

$$u_\kappa(\boldsymbol{x}) \;=\; \mathrm{P}(\Omega_\kappa|\boldsymbol{x}) \;\propto\; \mathrm{P}(\Omega_\kappa)\cdot\mathrm{P}(\boldsymbol{x}|\Omega_\kappa)$$

Sicherheit u/o Alternativen nicht aus $\hat{\kappa}(\boldsymbol{x})$ erschließbar

Trennfunktionen

Wie verlaufen die Grenzen zwischen je zwei Klassen?

$$h_{\kappa\lambda}(\boldsymbol{x}) \;=\; \log\frac{u_\kappa(\boldsymbol{x})}{u_\lambda(\boldsymbol{x})}$$

Partitionen

Welche Muster gehören zu welchen Klassen?

$$\hat{\kappa}(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\lambda=1..K} u_\lambda(\boldsymbol{x})$$

nicht aus $h_{\kappa\lambda}(\boldsymbol{x})$ wg. „Bermuda"-Syndrom

Naiver Bayesklassifikator

Idealisierende Annahme klassenweiser Merkmalunabhängigkeit

Lemma

Sind die Merkmale $x_1, \ldots, x_D$ der Muster jeder Klasse Ω_κ ***statistisch unabhängig****, so lautet die („**naive**") Bayes-Entscheidungsregel:*

$$\hat{\kappa}(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\lambda=1..K} \left(\mathrm{P}(\Omega_\lambda) \cdot \prod_{d=1}^{D} \mathrm{P}(x_d|\Omega_\lambda) \right)$$

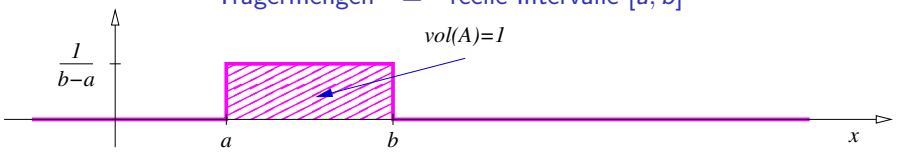
Bemerkung

Es sind $K \cdot D$ univariate Dichtefunktionen $f_{\lambda,d}(\cdot)$ aus den Lerndaten zu schätzen.

- Gleichverteilung (uniforme Dichte)
- Normalverteilung (Gaußdichte)
- Laplace- oder Cauchydichte (super-/subnormal)
- Parzen- oder Histogrammdichte (parameterfrei)

Univariate Gleichverteilungsdichten

Trägermengen $\hat{=}$ reelle Intervalle $[a, b]$



Definition

Ist $\mathcal{A} \subset \mathbb{R}^D$ eine D-dimensionale Punktmenge mit endlichem Hypervolumen $vol(\mathcal{A})$, so heißt

$$f_{\mathcal{A}}^{\text{unif}} : \boldsymbol{x} \mapsto \begin{cases} 1/vol(\mathcal{A}) & \boldsymbol{x} \in \mathcal{A} \\ 0 & \boldsymbol{x} \notin \mathcal{A} \end{cases}$$

die **Gleichverteilungs**- oder **uniforme Dichte** zum Träger(ereignis) $\mathcal{A}$.

Der univariate Spezialfall

$$f_{[a,b]}^{\text{unif}} : x \mapsto \begin{cases} 1/(b-a) & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}, \qquad a, b \in \mathbb{R}\,, \qquad a < b$$

heißt **uniforme Intervalldichte** über $[a, b]$.

Univariate Gleichverteilungsdichten

Maximum-Likelihood-Schätzung der Verteilungsparameter

Lemma

Für eine Lernprobe $\omega = \{z_1, \ldots, z_T\}$ nimmt die (unlogarithmierte) Maximum-Likelihood-Zielfunktion

$$\ell_{\text{ML}}(a, b) \;=\; f_{[a,b]}^{unif}(\omega) \;=\; \prod_{t=1}^{T} f_{[a,b]}^{unif}(z_t)$$

ihren maximalen Wert bei den Intervallgrenzen

$$a^* \overset{def}{=} \min_{t=1..T} z_t\,, \qquad b^* \overset{def}{=} \max_{t=1..T} z_t$$

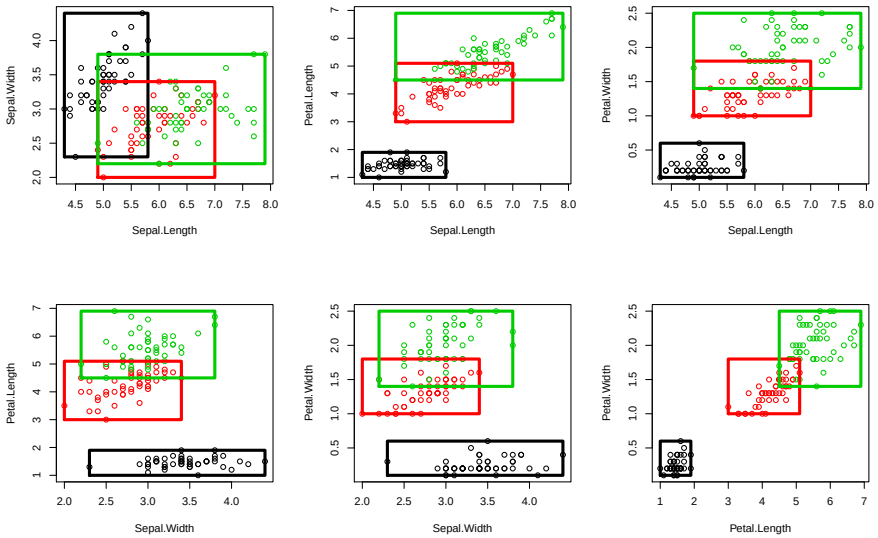
an.

Bemerkung

Es ist $\ell_{\text{ML}}(a, b) = 1 \,/\, (b-a)^T$, sofern alle z_t in $[a, b]$ liegen und Null sonst.

Univariate Gleichverteilungsdichten

Beispiel: Iris-Datensatz, 2D-Träger für alle (x_i, x_j)-Kombinationen



UNB — Entscheidungsregel

... für den naiven Bayesklassifikator mit Gleichverteilung

Prüfgröße

$$u_\kappa(\boldsymbol{x}) \;=\; \hat{\mathrm{P}}(\Omega_\kappa) \cdot \hat{f}_\kappa^{\text{unif}}(\boldsymbol{x}) \;=\; \frac{T_\kappa}{T} \cdot \frac{1}{\prod_d (b_{\kappa,d} - a_{\kappa,d})} \cdot \underbrace{\mathbf{I}_{\boldsymbol{a}_\kappa \le \boldsymbol{x} \le \boldsymbol{b}_\kappa}}_{\mathbf{I}_{\boldsymbol{x} \in \mathcal{H}_\kappa}}$$

Entscheidungsregel

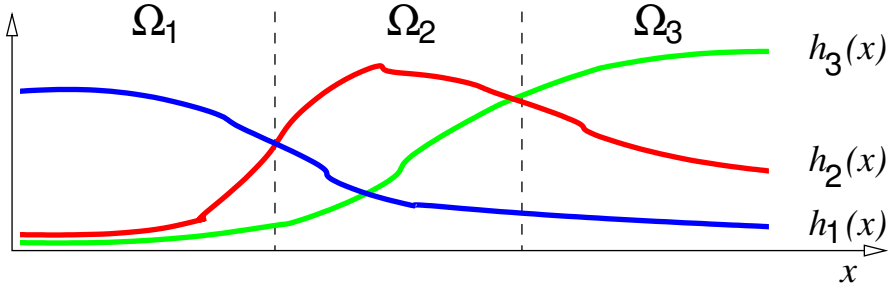
$$\hat{\kappa}(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\lambda=1..K} u_\kappa(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\lambda \mid \boldsymbol{x} \in \mathcal{H}_\lambda} \frac{T_\kappa}{vol(\mathcal{H}_\lambda)}$$

Bemerkung

- Die Muster außerhalb von $\bigcup_\lambda \mathcal{H}_\lambda$ können keiner Klasse zugeordnet werden.
- Klassen mit einem Ausreißer werden *extrem* benachteiligt ($vol(\mathcal{H}_\lambda) \to \infty$).
- Die UNB-Regel verallgemeinert miserabel gegenüber „neuen“ Mustern $\boldsymbol{x} \notin \omega$.

Verteilungsfreie Klassifikatoren

Direkte Approximation der a posteriori-Klassenwahrscheinlichkeit



Trenn- oder Diskriminantenfunktionen

Für jede Klasse ist die wahre Rückschlußwahrscheinlichkeit durch eine geeignete Modellfunktion anzunähern:

$$h_\kappa(\boldsymbol{x}) \approx \mathrm{P}(\Omega_\kappa|\boldsymbol{x}) = \frac{\mathrm{P}(\Omega_\kappa)\cdot\mathrm{P}(\boldsymbol{x}|\Omega_\kappa)}{\mathrm{P}(\boldsymbol{x})}$$

Approximationskriterium

In welchem Sinne ist $\mathrm{P}(\Omega_\kappa|\boldsymbol{x})$ durch $h_\kappa(\boldsymbol{x})$ anzunähern?

Minimaler Klassifikationsfehler

$$\#_{\kappa,\boldsymbol{x}}\left(h_\kappa(\boldsymbol{x}) \neq \max_\lambda h_\lambda(\boldsymbol{x})\right) \stackrel{!}{\to} \text{MIN}$$

Maximale Rückschlußwahrscheinlichkeit

logit-Modell oder *probit*-Modell; Optimierung mit linearen N.B.

$$\prod_{\kappa,\boldsymbol{x}} h_\kappa(\boldsymbol{x}) \stackrel{!}{\to} \text{MAX} \quad \text{wobei} \quad \sum_{\lambda=1}^{K} h_\lambda(\boldsymbol{x}) = 1 \quad (\forall \boldsymbol{x} \in \boldsymbol{\Omega})$$

Minimaler Rekonstruktionsfehler

Geometr. Abstand zur **idealen** Trennfunktion; Ausgleichsrechnung

$$\sum_{\kappa,\boldsymbol{x}} \|h_\lambda(\boldsymbol{x}) - \mathbf{I}_{\lambda=\kappa}\|^2 \stackrel{!}{\to} \text{MIN} \quad (\forall \lambda = 1..K)$$

Quadratmittelklassifikator

Die ideale Trennfunktion sagt 1/0 zur richtigen/falschen Klasse

Definition

Die erwartete quadratische Abweichung

$$\mathcal{E}_{\mathbb{X},\mathbb{K}}[\left\|\boldsymbol{h}(\mathbb{X}) - \boldsymbol{e}^{(\mathbb{K})}\right\|^2] = \sum_{\lambda=1}^{K}\sum_{\kappa=1}^{K}\int \mathrm{P}(\boldsymbol{x},\Omega_\kappa)\cdot(h_\lambda(\boldsymbol{x}) - \mathbf{I}_{\kappa=\lambda})^2\, d\boldsymbol{x}$$

heißt **Rekonstruktionsfehler** der idealen Trennfunktion einer Klassifikationaufgabe.

Bemerkungen

1. Die Approximation der idealen Trennfunktion ($\boldsymbol{h}(\boldsymbol{x})$ liefert 1 für die „richtige“ Klasse) fordert viel mehr als unbedingt nötig.
2. Trotzdem gilt, bei uneingeschränkter Optimierung: $h_\lambda(\boldsymbol{x}) = \mathrm{P}(\Omega_\lambda|\boldsymbol{x})$ besitzt den minimalen Rekonstruktionsfehler.
3. In praxi wird der Rekonstruktionsfehler auf einer klassifizierten Lernstichprobe ermittelt.

Linearer Quadratmittelklassifikator

Diskriminanten sind linear in den Merkmalvektorkomponenten

$$\boldsymbol{x} \in \mathbb{R}^D \quad \Rightarrow \quad \boldsymbol{h}(\boldsymbol{x}) \;=\; \begin{pmatrix} h_1 & = & \boldsymbol{a}_1^\top \boldsymbol{x} \\ h_2 & = & \boldsymbol{a}_2^\top \boldsymbol{x} \\ \dots & = & \dots \\ h_K & = & \boldsymbol{a}_K^\top \boldsymbol{x} \end{pmatrix} \quad \Rightarrow \quad \kappa^*(\boldsymbol{x}) = \operatorname*{argmax}_\lambda h_\lambda(\boldsymbol{x})$$

Lemma

Der Klassifikator mit den linearen Prüfgrößen

$$h_\lambda(\boldsymbol{x}) \;=\; \sum_{d=1}^{D} a_{\lambda,d} \cdot x_d \;, \quad \lambda = 1..K \qquad \textit{(kürzer: } \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{x}\textit{)}$$

besitzt auf der etikettierten Stichprobe $\{(\boldsymbol{x}_t, \kappa_t) \mid t = 1, 2, \dots, T\}$ *den (empirischen) Rekonstruktionsfehler*

$$\varepsilon(\boldsymbol{A}) \;=\; \sum_{\lambda=1}^{K} \varepsilon(\boldsymbol{a}_\lambda) \;=\; \sum_{\lambda=1}^{K} \sum_{t=1}^{T} (\boldsymbol{a}_\lambda^\top \boldsymbol{x}_t - \mathbf{I}_{\kappa_t = \lambda})^2 \;=\; \sum_{\lambda=1}^{K} \|\boldsymbol{X}\boldsymbol{a}_\lambda - \boldsymbol{y}_\lambda\|^2$$

mit der Datenmatrix $\boldsymbol{X}$ *und den Klassenindikatoren* $\boldsymbol{y}_\lambda$, $\lambda = 1..K$.

Beweis.

Der Gesamtfehler $\varepsilon(\boldsymbol{A})$ zerfällt in eine Summe ungekoppelter, klassenbezogener Fehlerkomponenten

$$\varepsilon(\boldsymbol{a}_\lambda) \;=\; \boldsymbol{a}_\lambda^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_\lambda - \boldsymbol{a}_\lambda^\top \boldsymbol{X}^\top \boldsymbol{y}_\lambda - \boldsymbol{y}_\lambda^\top \boldsymbol{X} \boldsymbol{a}_\lambda + \boldsymbol{y}_\lambda^\top \boldsymbol{y}_\lambda \;,$$

die separat optimiert werden können; dazu leiten wir partiell ab:

$$\nabla_{\boldsymbol{a}_\lambda} \varepsilon(\boldsymbol{a}_\lambda) \;=\; 2 \cdot \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_\lambda - 2 \cdot \boldsymbol{X}^\top \boldsymbol{y}_\lambda \;=\; 2 \cdot (T \cdot \boldsymbol{R}) \boldsymbol{a}_\lambda - 2 \cdot (T_\lambda \cdot \hat{\boldsymbol{\mu}}_\lambda)$$

Das Resultat setzen wir (komponentenweise) gleich Null, kürzen durch $2 \cdot T$ und erhalten die Gaußschen Normalengleichungen. □

Im Satz und im Beweis wurden folgende Hilfsgrößen verwendet:

$$\boldsymbol{R} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top \;, \qquad \boldsymbol{\mu}_\lambda = \frac{1}{T_\lambda} \sum_{t \mid \kappa_t = \lambda} \boldsymbol{x}_t \;, \qquad T_\lambda = \sum_{t \mid \kappa_t = \lambda} 1 \;, \qquad \hat{p}_\lambda = \frac{T_\lambda}{T}$$

Optimaler linearer Quadratmittelklassifikator

Gaußsches Normalengleichungssystem

Satz

Die Diskriminantenparameter $\boldsymbol{a}_\lambda$, $\lambda = 1..K$, *für den linearen Quadratmittelklassifikator mit minimalem Rekonstruktionsfehler* $\varepsilon(\boldsymbol{A})$ *ergeben sich aus den* ***Gaußschen Normalengleichungen***

$$\boldsymbol{R} \cdot \boldsymbol{a}_\lambda \;=\; \boldsymbol{m}_\lambda \;, \qquad \lambda = 1, \dots, K$$

mit den Stichprobenstatistiken

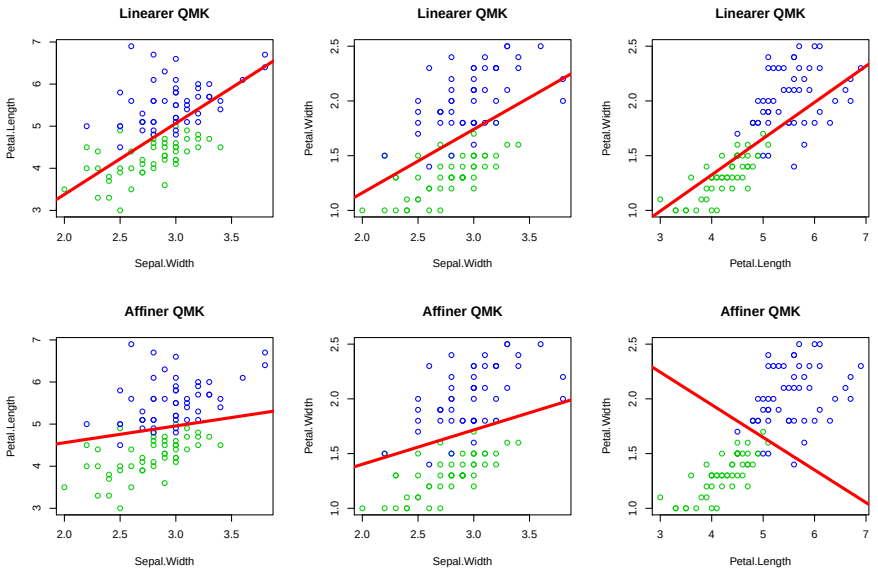
$$\boldsymbol{R} \;=\; \hat{\boldsymbol{S}} + \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^\top \quad \textit{und} \quad \boldsymbol{m}_\lambda \;=\; \hat{p}_\lambda \cdot \hat{\boldsymbol{\mu}}_\lambda \;.$$

Bemerkung

Ist die $(T \times T)$-Matrix $\boldsymbol{R}$ der zweiten Stichprobenmomente invertierbar, so gilt $\boldsymbol{A} = \boldsymbol{R}^{-1} \cdot \boldsymbol{M}$ mit $\boldsymbol{M} = (\boldsymbol{m}_1, \dots, \boldsymbol{m}_K)$.

Linearer vs. affiner Quadratmittelklassifikator

Beispiel: Iris-Datensatz, 2D-Träger für einige (x_i, x_j)-Kombinationen



Termlinearer Quadratmittelklassifikator

Diskriminanten sind linear in den Termkoeffizienten

$$\boldsymbol{x} \in \mathbb{R}^D \Rightarrow \phi(\boldsymbol{x}) = \underbrace{\begin{pmatrix} \phi_1(\boldsymbol{x}) \\ \phi_2(\boldsymbol{x}) \\ \vdots \\ \phi_L(\boldsymbol{x}) \end{pmatrix}}_{\boldsymbol{z} \in \mathbb{R}^L} \Rightarrow \boldsymbol{h}(\boldsymbol{z}) = \boldsymbol{g}^{\phi}(\boldsymbol{x}) = \underbrace{\begin{pmatrix} \boldsymbol{a}_1^\top \boldsymbol{z} \\ \boldsymbol{a}_2^\top \boldsymbol{z} \\ \vdots \\ \boldsymbol{a}_K^\top \boldsymbol{z} \end{pmatrix}}_{\boldsymbol{A}^\top \boldsymbol{z}} \Rightarrow \kappa^*(\boldsymbol{x})$$

⟨Algorithmus⟩

1. TERMEXPANSION
   Erweitere die Datenmatrix $\boldsymbol{X} \in \mathbb{R}^{T \times D}$ durch Termberechnung zu $\boldsymbol{X}^{\phi} \in \mathbb{R}^{T \times L}$.
2. STICHPROBENSTATISTIKEN
   Berechne die Stichprobenmomente $\hat{p}_\lambda$, $\hat{\boldsymbol{\mu}}_\lambda^{\phi}$ und $\boldsymbol{R}^{\phi}$ für alle Klassen $\lambda = 1..K$.
3. GAUßSCHE NORMALENGLEICHUNGEN
   Invertiere $\boldsymbol{R}^{\phi}$ und berechne alle Koeffizientenvektoren $\boldsymbol{a}_\lambda \in \mathbb{R}^L$.

⟨Algorithmus⟩

Quadratmittelklassifikator mit Polynomtermen

Riskantes Spiel mit dem Fluch der Dimension

- **Affine Terme** $L = D + 1$

$$a_{\lambda,0} + \sum_{i=1}^{D} a_{\lambda,i} \cdot x_i$$

- **Quadratische Terme** $L = {}^{(D+1)(D+2)}/_{2}$

$$a_{\lambda,0} + \sum_{i=1}^{D} a_{\lambda,i} \cdot x_i + \sum_{i=1}^{D} \sum_{j \geq i} b_{\lambda,i,j} \cdot x_i x_j$$

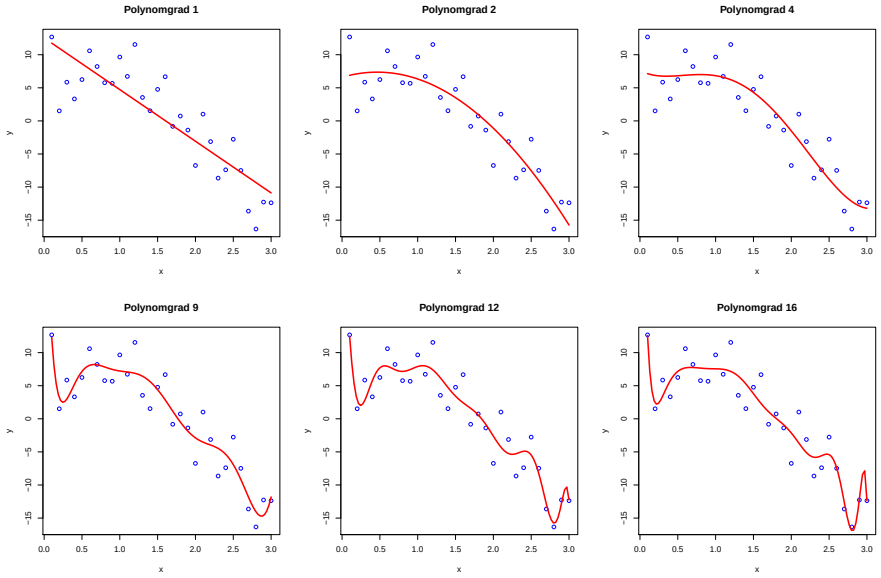
- **Kubische Terme** $L = {}^{(D+1)(D+2)(D+3)}/_{6}$

$$a_{\lambda,0} + \sum_{i=1}^{D} a_{\lambda,i} \cdot x_i + \sum_{i=1}^{D} \sum_{j \geq i} b_{\lambda,i,j} \cdot x_i x_j + \sum_{i=1}^{D} \sum_{j \geq i} \sum_{k \geq j} c_{\lambda,i,j,k} \cdot x_i x_j x_k$$

- **Polynomterme n-ten Grades** $L = \binom{D+n}{n}$

Überanpassungseffekt bei Ausgleichpolynomen

Weiß verrauschte Daten zur Kurve $y = 7 + 2x - 3x^2$



Nichtparametrische Dichteschätzung

Problem

Die Näherung der Verteilungen $\mathrm{P}(\boldsymbol{x}|\Omega_\kappa)$ durch parametrische Dichtefunktionen (z.B. Normal- oder Gleichverteilung) birgt das Risiko einer fehlerhaften Modellierung der Lerndaten ω_κ in sich.

Empirische Verteilung der Daten?

Die Auszählung der Ereignishäufigkeiten zur Formulierung einer kanonischen Verteilung

$$f^{\text{emp}}(\boldsymbol{x}|\omega_\kappa) \;\propto\; \begin{cases} 1 & \boldsymbol{x} \in \omega_\kappa \\ 0 & \boldsymbol{x} \notin \omega_\kappa \end{cases}$$

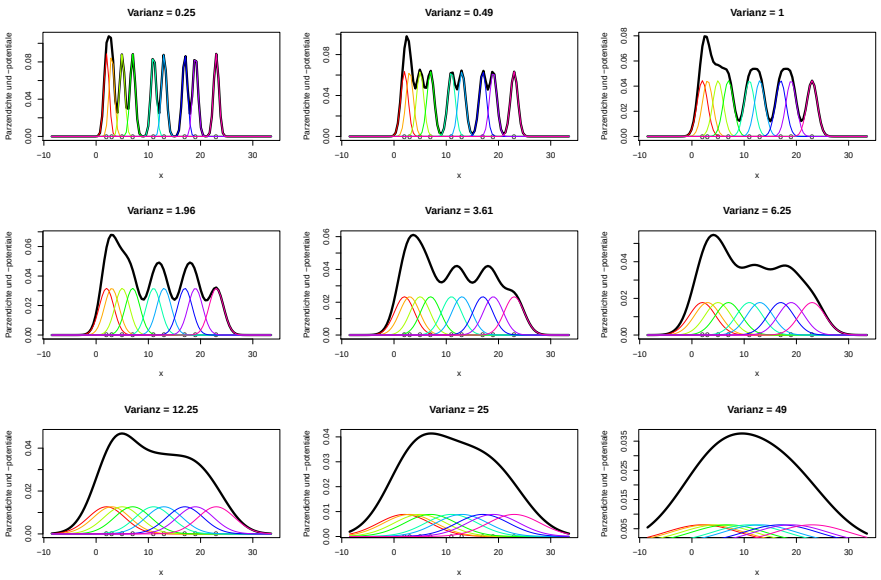
ist nur in diskreten Ereignisräumen opportun.

Lösung

Die (infinitesimal hohen und schmalen) Dirac-Gipfel der empirischen Datenverteilung werden durch geeignete Kern- oder Potentialfunktionen verflacht.

Parzen-Dichteschätzung

Potenzialfunktionen $\hat{=}$ Gaußkerne mit verschiedenen Streuungen σ^2



Parzen-Dichteschätzung

Potenzialfunktionen & Mittelung ihrer Verschiebungsinstanzen

Definition

Wir bezeichnen eine stetige Abbildung $g : \mathbb{R}^D \to \mathbb{R}_0^+$ als **Potenzialfunktion**, wenn sie flächennormiert ist und ihre Masse sich um den Ursprung konzentriert:

$$\int_{\mathbb{R}^D} g(\boldsymbol{\xi})\, d\boldsymbol{\xi} \;=\; 1 \qquad \text{und} \qquad \int_{\mathbb{R}^D} \|\boldsymbol{\xi}\|^2 \cdot g(\boldsymbol{\xi})\, d\boldsymbol{\xi} \;<\; \infty$$

Die Funktion

$$f_\omega^{\text{parzen}}(\boldsymbol{x}) \;=\; \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{x} - \boldsymbol{z}_t) \quad \text{bzw.} \quad f_{\omega,\sigma}^{\text{parzen}}(\boldsymbol{x}) \;=\; \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sigma^D} \cdot g\left(\frac{\boldsymbol{x} - \boldsymbol{z}_t}{\sigma}\right)$$

heißt **Parzendichte** der Datenprobe $\omega = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$ mit Potentialfunktion $g(\cdot)$ (und Konzentration $1/\sigma$).

Bemerkungen

1. $g(\cdot)$ ist Gaußglocke, Dreieck/Kegel, uniformes Rechteck/Hyperwürfel.
2. Mit $g(\boldsymbol{x})$ ist auch jede verschobene und skalierte Instanz normiert und $\boldsymbol{z}$-konzentriert.
3. Die oben definierte Summe ist offensichtlich nichtnegativ und flächennormiert, repräsentiert also eine Verteilungsdichte.

Parzen-Dichteschätzung

Theoretisches Resultat: Asymptotische Konvergenz

Satz

Die multivariate Zufallsvariable $\mathbb{X} \in \mathbb{R}^D$ sei gemäß der stetigen Dichte $f_\mathbb{X}$ verteilt und $\omega = \{\boldsymbol{z}_n | n \in \mathbb{N}\}$ sei eine Folge zufällig und unabhängig gezogener Stichprobenelemente.

Genügt die normierte und beschränkte Potentialfunktion $g : \mathbb{R}^D \to \mathbb{R}_0^+$ der Bedingung

$$\lim_{|\boldsymbol{x}|\to\infty} g(\boldsymbol{x}) \cdot \prod_{d=1}^{D} x_d \;=\; 0$$

und ist $[h_\nu]$ eine positive reelle Zahlenfolge mit den Eigenschaften

$$\lim_{\nu\to\infty} h_\nu^q = 0 \quad \textit{und} \quad \lim_{\nu\to\infty} \nu \cdot h_\nu^q = \infty\,, \qquad (\forall q \in \mathbb{N})$$

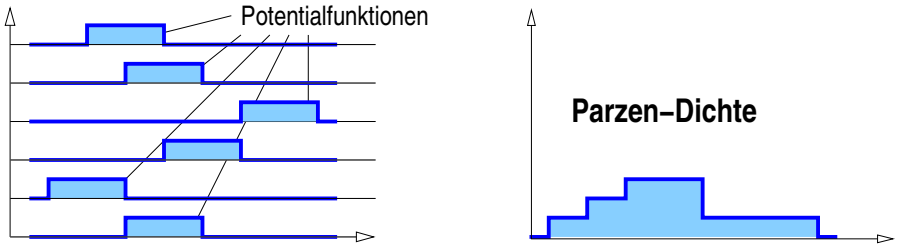
dann konvergiert die Parzen-Schätzung

$$\hat{\mathrm{P}}_N(\boldsymbol{x}) \;=\; \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h_N^D} \cdot g\left(\frac{\boldsymbol{x} - \boldsymbol{z}_n}{h_N}\right)$$

im quadratischen Mittel gegen die wahre Dichte $f_\mathbb{X}$.

Uniforme Parzenschätzung

Potenzialfunktion $\hat{=}$ Gleichverteilung auf Nullpunktumgebung



Definition

Eine Parzenschätzung mit

$$g(\boldsymbol{x}) = \begin{cases} 1/V & \boldsymbol{x} \in \mathcal{U} \\ 0 & \boldsymbol{x} \notin \mathcal{U} \end{cases}$$

für eine Nullpunktumgebung $\mathcal{U} \subset \mathbb{R}^D$ mit Hypervolumen $vol(\mathcal{U}) = V$ heißt **uniforme Parzenschätzung**.

Aufgabe Entscheidungstheorie Klassifikatortypen Statistisch Diskriminativ Parzen Induktion

Uniforme Parzenschätzung

UPS-Klassifikator — mit symmetrischer Nullumgebung

Klassendichte

$$\hat{\mathrm{P}}(\boldsymbol{x}|\Omega_\kappa) = \frac{m_\kappa^{(V)}(\boldsymbol{x})}{T_\kappa \cdot V}, \qquad m_\kappa^{(V)}(\boldsymbol{x}) \text{ ist die Anzahl der } \boldsymbol{z} \in \omega_\kappa \text{ mit } \boldsymbol{x} \in \mathcal{U}(\boldsymbol{z})$$

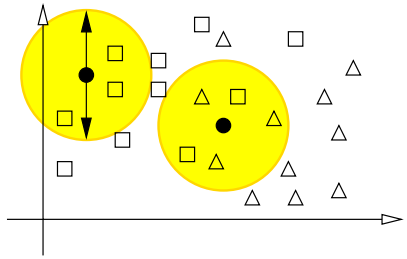
Prüfgröße

$$u_\kappa(\boldsymbol{x}) = \hat{p}_\kappa \cdot \hat{\mathrm{P}}(\boldsymbol{x}|\Omega_\kappa) = \frac{T_\kappa}{T} \cdot \frac{m_\kappa^{(V)}(\boldsymbol{x})}{T_\kappa \cdot V} = \frac{m_\kappa^{(V)}(\boldsymbol{x})}{T \cdot V}$$

Entscheidungsregel

$$\delta(\boldsymbol{x}) = \operatorname*{argmax}_\kappa m_\kappa^{(V)}(\boldsymbol{x})$$

(maximale Anzahl der ω_κ-Befunde in $\mathcal{U}(\boldsymbol{x})$)



Aufgabe Entscheidungstheorie Klassifikatortypen Statistisch Diskriminativ Parzen Induktion

Histogramm-Klassifikator

UPS-ähnlicher Klassifikator — und mit ganz ähnlichen Problemen

⟨Algorithmus⟩

1. PARZELLIERUNG
   Zerlege Merkmalraum $\mathbb{R}^D$ in Hyperquader mit Volumina $V = \Delta^D$.
2. TREFFERQUOTEN SPEICHERN
   Je Quader $\mathcal{Q}$ setze $m_\kappa(\mathcal{Q}) = \#(\omega_\kappa \cap \mathcal{Q})$.
3. ENTSCHEIDUNGSREGEL
   Klassifiziere $\boldsymbol{x}$ nach maximalem $m_\kappa(\mathcal{Q}_{\boldsymbol{x}})$ in dem Quader mit $\boldsymbol{x} \in \mathcal{Q}_{\boldsymbol{x}}$.

⟨Algorithmus⟩

Fluch der Dimension

Fast alle Zellen ($\mathcal{Q}_{\boldsymbol{x}}$, $\mathcal{U}(\boldsymbol{x})$) sind leer!

Beispiel

- Merkmalraum ist $\mathbb{R}^{20}$
- Stichprobenumfang $N = |\omega| = 10^6$ Muster
- Ca. 10 Intervalle je Koordinate

Von den 10^{20} *Zellen sind mindestens* $10^{20} - 10^6$ *leer ...*

Aufgabe Entscheidungstheorie Klassifikatortypen Statistisch Diskriminativ Parzen Induktion

Hypersphären-Schätzung und -klassifikator

Wie UPS — aber Befundzahl fixieren und Volumen variieren

Klassendichte

$$\hat{\mathrm{P}}(\boldsymbol{x}|\Omega_\kappa) = \frac{m}{T_\kappa \cdot V_\kappa^{(m)}(\boldsymbol{x})}, \qquad \text{mit } V_\kappa^{(m)}(\boldsymbol{x}) = \min_\rho \{vol(\mathcal{U}_\rho(\boldsymbol{x})) \mid |\mathcal{U}_\rho(\boldsymbol{x}) \cap \omega_\kappa| \geq m\}$$

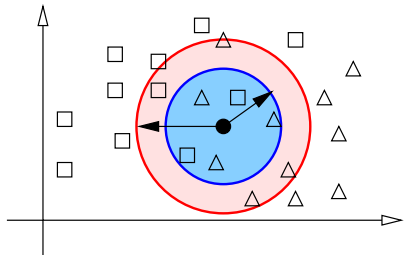
Prüfgröße

$$u_\kappa(\boldsymbol{x}) = \hat{p}_\kappa \cdot \hat{\mathrm{P}}(\boldsymbol{x}|\Omega_\kappa) = \frac{T_\kappa}{T} \cdot \frac{m}{T_\kappa \cdot V_\kappa^{(m)}(\boldsymbol{x})} = \frac{m}{T \cdot V_\kappa^{(m)}(\boldsymbol{x})}$$

Entscheidungsregel

$$\delta(\boldsymbol{x}) = \operatorname*{argmin}_\kappa V_\kappa^{(m)}(\boldsymbol{x})$$

(ω_κ versammelt m Muster auf kleinstem Raum um $\boldsymbol{x}$)



k-Nächste-Nachbarn-Regel

Wie HSK — aber gesucht sind Befunde *beliebiger* Klasse

Prüfgröße

UPS der Rückschlußwahrscheinlichkeiten

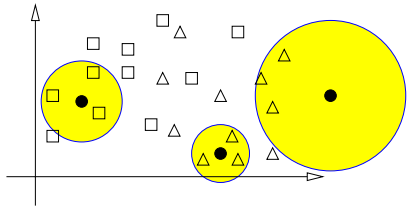
$$\hat{\mathrm{P}}(\Omega_\kappa|\boldsymbol{x}) \;=\; \frac{\hat{\mathrm{P}}(\Omega_\kappa)\cdot\hat{\mathrm{P}}(\boldsymbol{x}|\Omega_\kappa)}{\hat{\mathrm{P}}(\boldsymbol{x})} \;=\; \frac{\frac{T_\kappa}{T}\cdot\frac{m_\kappa^{(k)}(\boldsymbol{x})}{T_\kappa\cdot V}}{\frac{k}{T\cdot V}} \;=\; \frac{m_\kappa^{(k)}(\boldsymbol{x})}{k}$$

Unter den k nächsten Nachbarn von $\boldsymbol{x}$ in ω (Hypervolumen $= V$) seien jeweils genau $m_\kappa^{(k)}(\boldsymbol{x})$ aus der Klasse κ gewesen.

Entscheidungsregel

$$\delta(\boldsymbol{x}) \;=\; \operatorname*{argmax}_\kappa m_\kappa^{(k)}(\boldsymbol{x})$$

(die Mehrheitsfraktion unter den k nächsten $\boldsymbol{x}$-Nachbarn)



Nächster-Nachbar-Regel

NNR $\hat{=}$ Spezialfall von HSK ($m = 1$) und auch von kNN ($k = 1$)

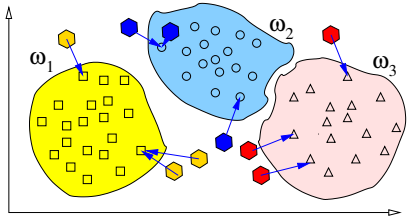
Prüfgröße

$$\underbrace{u_\kappa(\boldsymbol{x}) \;=\; \frac{1}{T\cdot V_\kappa^{(1)}(\boldsymbol{x})} \quad \text{bzw.} \quad u_\kappa(\boldsymbol{x}) \;=\; \frac{m_\kappa^{(1)}(\boldsymbol{x})}{1}}_{u_\kappa(\boldsymbol{x}) \;=\; \frac{1}{T\cdot\rho_\kappa^D}\,,\quad \rho_\kappa = \min\limits_{\boldsymbol{z}\in\omega_\kappa}\|\boldsymbol{x}-\boldsymbol{z}\|}$$

Entscheidungsregel

$$\delta(\boldsymbol{x}) \;=\; \operatorname*{argmin}_\kappa \min_{\boldsymbol{z}\in\omega_\kappa}\|\boldsymbol{x}-\boldsymbol{z}\|$$

(Klassenindex des nächsten Nachbarn von $\boldsymbol{x}$ in ω)



Nichtparametrische Klassifikatoren

Vorteile

- Keine anfechtbaren Vorannahmen über die Verteilung oder die Trennflächen der Musterklassen
- Keine aufwändige Lernprozedur

Nachteile

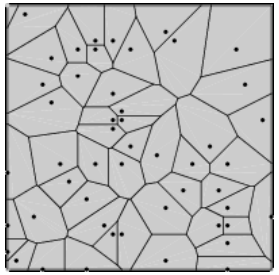
- Irrsinnig hoher Klassifikationsaufwand: $\mathrm{O}(T\cdot D)$ je Muster
- Irrsinnig hoher Speicherbedarf: $\mathrm{O}(T\cdot D)$ für ω

Offene Fragen

1. Welches V bzw. Δ für UPS- und Histogramm-Klassifikatoren?
2. Welche Befundraten m bzw. k für Hypersphären- und kNN-Klassifikatoren?
3. Alternative Distanzmaße für Hypersphären- und kNN-Klassifikatoren?

Voronoizellen

Nächster-Nachbar-Regel mit euklidischem Abstand



Voronoipartition der Ebene

Definition

Für eine Teilmenge $\omega = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\}$ des Raumes $\mathbb{R}^D$ heißen die Mengen

$$\mathcal{V}_t \;\stackrel{\text{def}}{=}\; \{\boldsymbol{z} \mid \|\boldsymbol{z}-\boldsymbol{x}_t\|_2 \le \|\boldsymbol{z}-\boldsymbol{x}_s\|_2,\; s\neq t\}$$

die **Voronoizellen** von ω.

Bemerkungen

1. Die Voronoizellen $\mathcal{V}_t$, $t = 1..T$, besitzen paarweise leeren Durchschnitt oder einen Durchschnitt vom Volumen Null; das Mengensystem wird deshalb oft als **Voronoipartition** bezeichnet.
2. Nur im Inneren der Zellen ist die 1NN-Regel eindeutig.

Voronoipartition und Delauneytriangulierung

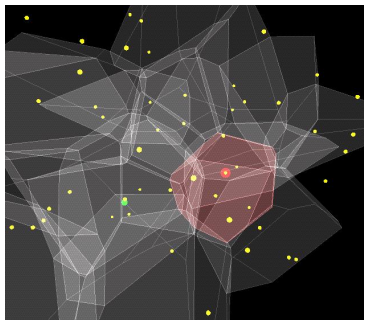
Topologische Repräsentation für schnelle Suchverfahren mehr Information

Definition

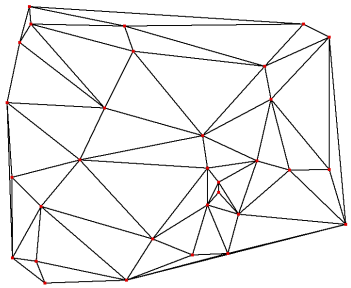
Für eine Teilmenge $\omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ des Raumes $\mathbb{R}^D$ bezeichnen wir den ungerichteten Graphen $(\mathcal{X}, \mathcal{E})$ mit $\mathcal{X} = \{1, \ldots, T\}$ und

$$\{s,t\} \in \mathcal{E} \quad \Leftrightarrow \quad \mathcal{V}_s \cap \mathcal{V}_t \neq \varnothing$$

als **Delauneytriangulierung** der Punktmenge.



Voronoizellen im $\mathbb{R}^3$ Wolfram



Delauneygraph

Klassifikator versus Klassifikationsverfahren

Konkreter Klassifikator

Entscheidungsregel im Raum $\boldsymbol{\Omega}$ für K Klassen:

$$\delta : \left\{ \begin{array}{ccc} \boldsymbol{\Omega} & \rightarrow & \{0,1\}^K \text{ oder } [0,1]^K \\ \boldsymbol{x} & \mapsto & (\delta_1(\boldsymbol{x}), \ldots, \delta_K(\boldsymbol{x})) \end{array} \right. , \qquad \sum_{\lambda} \delta_\lambda(\boldsymbol{x}) = 1$$

Etikettierte Stichprobe

K-dimensionales Feld von Mustersequenzen (Multimengen)

$$\omega : \ \{1, 2, \ldots, K\} \ \rightarrow \ \boldsymbol{\Omega}^*$$

Abstraktes Klassifikationsverfahren

Aus einer endlichen, etikettierten Lernstichprobe wird eine Entscheidungsregel für K Klassen gelernt:

$$\mathfrak{A} : \left\{ \begin{array}{ccc} (\boldsymbol{\Omega}^*)^K & \rightarrow & ([0,1]^K)^{\Omega} \\ \omega = [\omega_\kappa] & \mapsto & \delta_{\omega,\mathfrak{A}} \end{array} \right.$$

Reklassifikationsfehlerrate

Lernstichprobe $\hat{=}$ Teststichprobe

$$\boldsymbol{\Omega} \supset \omega \xrightarrow{\text{Lernen}} \delta_\omega \xrightarrow{\text{Testen}} \hat{p}_\varepsilon(\omega|\delta_\omega)$$

Definition

Für ein Klassifikationsverfahren $\mathfrak{A}$ bezeichnen wir die relative Häufigkeit

$$\hat{p}_\varepsilon(\omega|\delta_{\omega,\mathfrak{A}})$$

von Fehlklassifikationen der gelernten Entscheidungsregel auf ihren eigenen Lerndaten $\omega = (\omega_1, \ldots, \omega_K)$ als **Reklassifikationsfehlerrate**.

Bemerkungen

1. Die Reklassifikationsrate ist nicht eindeutig für das Verfahren $\mathfrak{A}$.
2. Die Reklassifikationsrate unterschätzt den Fehler bei Anwendung der Entscheidungsregel auf *neue* Muster.
3. Der Trend zur *optimistischen* Bewertung ist besonders kraß, wenn ω geringen Umfang besitzt.

Reklassifikationsfehlerrate

Iris-Datensatz · 3 Klassen · je 50 Muster des $\mathbb{R}^4$

Uniform-NB

5.3%	set	ver	vir
set	50	0	0
ver	0	50	0
vir	0	8	42

Affin-QMK

15.3%	set	ver	vir
set	50	0	0
ver	0	34	16
vir	0	7	43

1NN-Regel

0.0%	set	ver	vir
set	50	0	0
ver	0	50	0
vir	0	0	50

Beispiele

Der Fehler der 1NN-Regel beträgt praktisch Null.
Die UNB-Regel patzt u.U. in den Überlappungsgebieten.
Der QMK tendiert für hohen Polynomgrad zum Fehler Null.

Neuklassifikationsfehlerrate

Held-out: Lernstichprobe ! = Teststichprobe

$$\Omega \supset \omega_L \xrightarrow{\text{Lernen}} \delta_{\omega_L} \xrightarrow{\text{Testen}} \hat{p}_\varepsilon(\omega_T|\delta_{\omega_L})$$

Definition

Für ein Klassifikationsverfahren $\mathfrak{A}$ bezeichnen wir die relative Häufigkeit

$$\hat{p}_\varepsilon(\omega_T|\delta_{\omega_L,\mathfrak{A}})$$

von Fehlklassifikationen der gelernten Entscheidungsregel auf einer zu den Lerndaten ω_L disjunkten Testdatenprobe ω_T als **Held-out-Fehlerrate**.

Bemerkungen

1. Die Held-out-Fehlerrate auf ω_T ist ein Näherungswert der Fehlerrate, die sich bei Anwendung von $\delta_{\omega_L,\mathfrak{A}}$ auf *neue* Muster ergibt.
2. Die Held-out-Fehlerrate von $\delta_{\omega_L,\mathfrak{A}}$ unterschätzt die Leistungsfähigkeit des Verfahrens $\mathfrak{A}$, da die Entscheidungsregel nur bezüglich einer endlichen Probe ω_L optimiert wurde.
3. Der Trend zur *pessimistischen* Bewertung ist besonders kraß, wenn ω_L geringen Umfang besitzt.

Held-out-Fehlerrate

Iris-Datensatz · 50/100 Muster zum Lernen/Testen

Uniform-NB

16%	set	ver	vir
set	34	0	0
ver	5	28	0
vir	10	1	22

Affin-QMK

21%	set	ver	vir
set	34	0	0
ver	0	16	17
vir	0	4	29

1NN-Regel

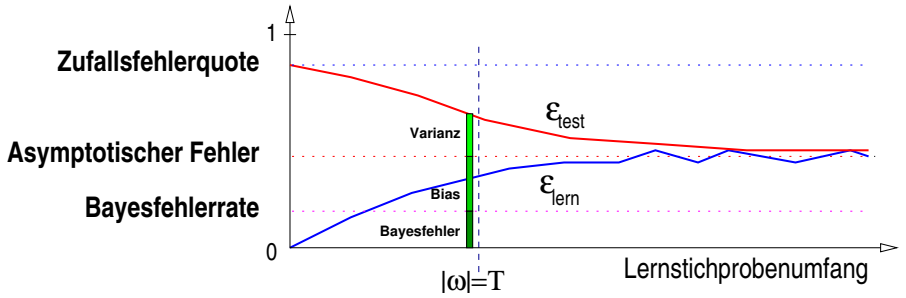
4%	set	ver	vir
set	34	0	0
ver	0	31	2
vir	0	2	31

Beispiele

Der Fehler der 1NN-Regel beträgt 4% statt 0%.
Die UNB-Regel verdreifacht ihren Fehler.
Beim QMK erhöht sich der Fehler um ein Drittel.

Die Diskrepanz der Fehlerraten korrespondiert mit der Anzahl der Freiheitsgrade.

Fehlerrate, Überanpassung und Unteranpassung



- **Bayesfehler** — was theoretisch herauszuholen ist (Bayesregel)
- **Grenzfehler** — das Verfahren $\mathfrak{A}$ mit infiniter Lernprobe
- **Zufallsfehler** — die „blinde“ Entscheidung: häufigste Klasse
- **Induktiver Bias** — schuld ist das unzureichende Datenmodell
- **Varianz** — schuld ist unsere begrenzte Lernstichprobe

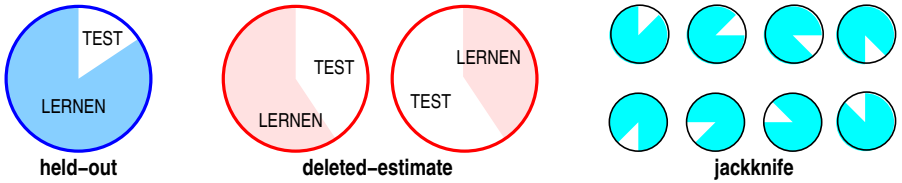
Rotationsverfahren

Robuste und ökonomische Klassifikatorentwurfstechnik

Fakt

Der Stichprobenumfang $|\omega^{lern}|$ ist verantwortlich für die Güte des angelernten Klassifikators.

Der Stichprobenumfang $|\omega^{test}|$ ist verantwortlich für die Genauigkeit seiner geschätzten Fehlerrate.



Rotationsverfahren

Benutze (fast) alle Muster zum Lernen UND zum Testen!

⟨Algorithmus⟩

1. ZERLEGE die Daten DISJUNKT: $\omega = \omega^{(1)} \cup \omega^{(2)} \cup \ldots \cup \omega^{(R)}$.
2. LERNE die Klassifikatoren $\delta^{(r)} = \delta_{\omega^{(r)}}$, $\omega^{(r)} = \omega \setminus \omega^r$, $r = 1..R$.
3. KLASSIFIZIERE alle Muster $\boldsymbol{x} \in \omega^r$ mit der Regel $\delta^{(r)}$.
4. KUMULIERE den absoluten und relativen Gesamtfehler.

⟨mhtirogIA⟩

Bemerkungen

1. *Deleted-estimate* ($R = 2$)
   Jackknife (R-fache Kreuzvalidierung)
   Leave-one-out ($R = |\omega|$).
2. Datenpartitionierung zufällig oder systematisch?
   (Klassenzugehörigkeit, Doubletten, Zeitreihen)

Rotations-Fehlerrate

Iris-Datensatz · 10-fache Kreuzvalidierung

Uniform-NB

11.3%	set	ver	vir
set	50	0	0
ver	3	43	4
vir	3	7	40

Affin-QMK

16%	set	ver	vir
set	49	1	0
ver	0	35	15
vir	0	8	42

1NN-Regel

4%	set	ver	vir
set	50	0	0
ver	0	47	3
vir	0	3	47

Beispiele

Die beiden parametrischen Klassifikatoren profitieren vom erhöhten Umfang der Lerndatenprobe (135 statt 50 Muster).

Die 1NN-Regel war offenbar bereits bei 50 Lernmustern in den Sättigungsbereich gelangt.

Genauigkeit der Fehlerschätzung

Wie viele Testmuster werden für eine seriöse Fehlervoraussage benötigt?

Fakt

Ein Klassifikationstest entspricht dem Ziehen aus einer Urne mit einem (unbekannten) Anteil p_ε von „Nieten“.

Lösung

Für einen Klassifikator δ mit wahrer Fehlerrate p_ε gehorcht die diskrete Zufallsvariable

$$\mathbb{M} \;=\; \textit{Anzahl falschklassifizierter unter } N \textit{ Mustern}$$

der ***Binomialverteilung****:*

$$\mathrm{P}(\mathbb{M} = n_f) \;=\; \mathcal{B}(n_f \mid p_\varepsilon, N) \;=\; \binom{N}{n_f} \cdot p_\varepsilon^{n_f} \cdot (1 - p_\varepsilon)^{N - n_f}$$

Genauigkeit der Fehlerschätzung

Große Teststichprobe ➪ Fehleranzahl ist normalverteilt

Satz (Binomialer Grenzwertsatz)

Ist die Zufallsvariable $\mathbb{X}_N$ für jedes $N \in \mathbb{N}$ binomialverteilt gemäß $\mathcal{B}(\cdot|N,p)$, so gilt die asymptotische Näherungsformel

$$\lim_{N\to\infty} \mathrm{P}\left(\frac{\mathbb{X}_N - \mathcal{E}[\mathbb{X}_N]}{\sqrt{\mathrm{Var}[\mathbb{X}_N]}} \leq x\right) \;=\; \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\, d\xi$$

mit den Mittelwerten $\mu = \mathcal{E}[\mathbb{X}_N] = N \cdot p$ und Varianzen $\sigma^2 = \mathrm{Var}[\mathbb{X}_N] = N \cdot p \cdot (1-p)$ der Binomialverteilung.

Beispiel

Für den Kreuzvalidierungstest des QMK auf den Iris-Daten gilt:

$$N = 150\;, \quad \hat{p}_\varepsilon = 0.16\;, \quad \mu = 24\;, \quad \sigma^2 = 20.16\;, \quad \sigma \approx 4.5$$

Folglich liegt mit 95% Sicherheit die Zahl der Nieten im Intervall $[15, 33]$, die wahre Fehlerrate also zwischen 10% und 22% — ein katastrophales Paradebeispiel indiskutabler Evaluierungsbedingungen!

Analytische Aussagen zur Bayesfehlerrate?

Satz (Fehlerrate der NN-Regel)

Ist p_ε^{Bayes} die Bayesfehlerrate eines Klassifikationsproblems $[\Omega_\kappa]$ und p_ε^{NN} die asymptotische Fehlerrate ($|\omega| \to \infty$) des Nächster-Nachbar-Klassifikators, so gilt

$$p_\varepsilon^{Bayes} \;\leq\; p_\varepsilon^{NN} \;\leq\; p_\varepsilon^{Bayes} \cdot \left(2 - \frac{K}{K-1} \cdot p_\varepsilon^{Bayes}\right)\;.$$

Für hinreichend kleine Fehlerraten folgt daraus

$$p_\varepsilon^{NN}/2 \;\leq\; p_\varepsilon^{Bayes} \;\leq\; p_\varepsilon^{NN}\;.$$

Bemerkungen

1. Die simple L^1O-Evaluierung eines Datensatzes liefert eine gute Schätzung der *nicht-asymptotischen* Fehlerrate $p_\varepsilon^{\mathrm{NN}}(\omega)$.
2. Für sehr große Proben ω können wir schließen, daß die Bayesfehlerrate oberhalb $\hat{p}_\varepsilon^{\mathrm{NN}}(\omega)\,/\,2$ liegt.

➪ Wackeliges Kriterium für *„Klassifikatorentwicklung erfolgreich abgeschlossen“*

Genauigkeit der Fehlerschätzung

Signifikanzniveau & Konfidenzintervall

Faustregel

Wenn wir auf ein *Signifikanzniveau* von $\mu \pm C\sigma$ orientieren und gegenüber der unbekannten Fehlerrate p_ε einen relativen Fehler von $\Delta > 0$ unterschreiten möchten, so werden mindestens

$$N \;=\; {}^{(1-p_\varepsilon)}\!/_{p_\varepsilon} \cdot {}^{C^2}\!/_{\Delta^2}$$

Testmuster benötigt

Bemerkungen

1. Zur Signifikanzschwelle $C = 3$ (99%) und relativer Abweichung $\Delta = 10\%$ sind 2700, 17000 oder 90000 Muster gefordert für Fehlerraten um 25%, 5% oder 1%.
2. Die Übertragung der Fehlerschranken für geschätzte Rate bezüglich wahrer Rate auf Fehlerschranken für die wahre Rate bezüglich des Schätzwerts ist natürlich strenggenommen ein (unzulässiger) *Abduktionsschluß*.

Beweis.

Alle Abschätzungen der Fehlerraten (0/1-Risiko) analysieren die punktweisen Fehlerwahrscheinlichkeiten $p_\varepsilon(\boldsymbol{x})$ für Bayesregel und 1NN-Regel und bilden abschließend die Erwartungswerte.

Es bezeichne $p_\lambda(\boldsymbol{x})$ die wahre a posteriori Wahrscheinlichkeit, daß Muster $\boldsymbol{x}$ zur Klasse Ω_λ gehört.

Die Bayesregel klassifiziert $\boldsymbol{x}$ nach maximalem $p_\lambda(\boldsymbol{x})$; das Risiko beträgt also

$$p_\varepsilon^{\mathrm{Bayes}}(\boldsymbol{x}) \;=\; 1 - \max_\lambda p_\lambda(\boldsymbol{x})\;.$$

Die asymptotische 1NN-Regel findet zu $\boldsymbol{x}$ einen nächsten Nachbarn $\tilde{\boldsymbol{x}} \in \omega$, der wegen $|\omega| \to \infty$ infinitesimal nahe bei $\boldsymbol{x}$ liegt. Folglich besitzt $\tilde{\boldsymbol{x}}$ dieselbe a posteriori Verteilung wie $\boldsymbol{x}$ selbt und wird von 1NN gemäß $p_\lambda(\boldsymbol{x})$ klassifiziert. Das Risiko beträgt also

$$p_\varepsilon^{\mathrm{NN}}(\boldsymbol{x}) \;=\; \sum_\lambda p_\lambda(\boldsymbol{x}) \cdot (1 - p_\lambda(\boldsymbol{x})) \;=\; 1 - \sum_\lambda p_\lambda^2(\boldsymbol{x})\;.$$

Die punktweise Abschätzung von $p_\varepsilon^{\mathrm{NN}}(\boldsymbol{x})$ nach unten und oben ergibt sich durch mehr oder weniger langwierige Standardumformungen ... □

Univariate Gleichverteilungsdichten

Leave-One-Out-Schätzung der Verteilungsparameter

$$a^* = \min_{t=1..T} z_t - \alpha\,, \qquad b^* = \max_{t=1..T} z_t + \beta$$

Lemma

Für eine Lernprobe $\omega = \{z_1, \ldots, z_T\}$ nimmt die (unlogarithmierte) Leave-One-Out-Zielfunktion

$$\ell_{\mathrm{L^1O}}(\alpha, \beta) = \prod_{t=1}^{T} f^{unif}_{[a^{(t)}, b^{(t)}]}(z_t)$$

mit den Intervallgrenzschätzern

$$a^{(t)} \stackrel{def}{=} \min_{s \neq t} z_s - \alpha\,, \qquad b^{(t)} \stackrel{def}{=} \max_{s \neq t} z_s + \beta$$

ihren maximalen Wert bei den Dilatationsparametern

$$\alpha^* \stackrel{def}{=} z_{(2)} - z_{(1)}\,, \qquad \beta^* \stackrel{def}{=} z_{(T)} - z_{(T-1)}$$

an, d.h., $a^* = 2z_{(1)} - z_{(2)}$ und $b^* = 2z_{(T)} - z_{(T-1)}$.

Axiomatische Wahrscheinlichkeitstheorie

Zufallsexperiment · Elementarereignisse · Ereignisse

Definition

Es sei $\mho$ der Raum aller möglichen Ausgänge eines „Zufallsexperiments“. Ein Mengensystem $\mathfrak{E} \subseteq \mathfrak{P}\mho$ heißt **Ereignisraum** über $\mho$, wenn es die folgenden Eigenschaften einer σ-**Algebra** besitzt:

(1) $\mho \in \mathfrak{E}$

(2) $A \in \mathfrak{E} \;\Rightarrow\; A^c = \mho \backslash A \in \mathfrak{E}$

(3) $A_n \in \mathfrak{E}\ (\forall n \in \mathbb{N}) \;\Rightarrow\; \bigcup_{n \in \mathbb{N}} A_n \in \mathfrak{E}$

Die Elemente von $\mathfrak{E}$ heißen **Ereignisse**, die Elemente von $\mho$ heißen **Elementarereignisse**. Wir bezeichnen ferner $\mho, \varnothing \in \mathfrak{E}$ als das **sichere** bzw. das **unmögliche** Ereignis.

Axiomatische Wahrscheinlichkeitstheorie

Eigenschaften und Beispiele von Ereignisräumen

Lemma

Ein Ereignisraum $\mathfrak{E}$ über $\mho$ ist insbesondere eine Boolesche Algebra:

(4) $\varnothing \in \mathfrak{E}$

(5) $A_1, A_2 \in \mathfrak{E} \;\Rightarrow\; A_1 \cup A_2 \in \mathfrak{E}$

(6) $A_1, A_2 \in \mathfrak{E} \;\Rightarrow\; A_1 \cap A_2 \in \mathfrak{E}$

Beispiele

1. $\mho = \{1, 2, 3, 4, 5, 6\}$ (Würfel), $\mathfrak{E} = \mathfrak{P}\mho$, $|\mathfrak{E}| = 2^6 = 64$
2. $\mho = \{(a, b, c) \mid a, b, c \in \{\mathrm{K}, \mathrm{Z}\}\}$ (drei Münzen), $|\mathfrak{P}\mho| = 2^8$
3. $\mho = \mathbb{N}_0$ (Verkehrstote Deutschland 1984)
   $\mathfrak{E} = \{A_k \mid k \in \mathbb{N}_0\}$ mit $A_k \stackrel{\text{def}}{=} \{k\}$ oder $A_k \stackrel{\text{def}}{=} \{1, \ldots, k\}$
4. $\mho = \mathbb{R}_0^+$ (Lebensdauer einer Glühbirne)
   $\mathfrak{E} = \{A_{r,\delta} \mid r, \delta \in \mathbb{R}_0^+\}$ mit $A_{s}r,\delta \stackrel{\text{def}}{=} \{x \mid r \leq x \leq r + \delta\}$
5. $\mho = \mathbb{R}^5$ (Ertrag von fünf Weizenarten), z.B. $A_{3,1} = \{\boldsymbol{x} \mid x_3 \geq 2x_1\}$

Axiomatische Wahrscheinlichkeitstheorie

Kolmogorov-Axiome für Wahrscheinlichkeitsräume

Definition (Kolmogorov)

Ist $\mathfrak{E}$ ein Ereignisraum über $\mathfrak{V}$, so heißt die Abbildung

$$\mathrm{P} : \mathfrak{E} \to \mathbb{R}$$

eine **Wahrscheinlichkeitsfunktion** über $\mathfrak{V}$ falls gilt:

$$\begin{array}{ll} (1) & \mathrm{P}(A) \geq 0 \quad \text{für alle } A \in \mathfrak{E} \\ (2) & \mathrm{P}(\mathfrak{V}) = 1 \\ (3) & \mathrm{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathrm{P}(A_n) \end{array}$$

Dabei seien die A_n, A_m paarweise disjunkt.

Die Eigenschaft (3) heißt σ-**Additivität** von P. Das Tripel $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ heißt dann **Wahrscheinlichkeitsraum** über $\mathfrak{V}$.

Axiomatische Wahrscheinlichkeitstheorie

Gleichungen und Ungleichungen für Wahrscheinlichkeitsräume

Satz

In einem Wahrscheinlichkeitsraum $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ gelten für alle $A, B, A_n \in \mathfrak{E}$ die Eigenschaften:

$$\begin{array}{ll} (1) & \mathrm{P} : \mathfrak{E} \to [0,1] \\ (2) & \mathrm{P}(\varnothing) = 0 \\ (3) & \mathrm{P}(A^c) = 1 - \mathrm{P}(A) \\ (4) & \mathrm{P}(A) = \mathrm{P}(A,B) + \mathrm{P}(AB^c) \\ (5) & \mathrm{P}(A\backslash B) = \mathrm{P}(AB^c) = \mathrm{P}(A) - \mathrm{P}(A,B) \\ (6) & \mathrm{P}(A\cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A,B) \\ (7) & A \subseteq B \Rightarrow \mathrm{P}(A) \leq \mathrm{P}(B) \\ (8) & \mathrm{P}(\bigcup_{n\in\mathbb{N}} A_n) \leq \sum_{n\in\mathbb{N}} \mathrm{P}(A_n) \end{array}$$

Dabei sei die Kurzschreibweise $A, B \stackrel{def}{=} A\cap B$ vereinbart.

Axiomatische Wahrscheinlichkeitstheorie

Endliche Wahrscheinlichkeitsräume · Gleichverteilung

Bemerkung

Ein Wahrscheinlichkeitsraum $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ mit endlicher Menge $\mathfrak{V} = \{\varepsilon_1, \ldots, \varepsilon_N\}$ von Elementarereignissen ist vollständig durch die Wahrscheinlichkeitsmassen $p_j = \mathrm{P}(\{\varepsilon_j\})$, $j = 1, \ldots, N$ charakterisiert, denn für jedes Ereignis $A \in \mathfrak{E} = \mathfrak{P}\mathfrak{V}$ gilt dann:

$$\mathrm{P}(A) = \sum_{\varepsilon\in A} \mathrm{P}(\{\varepsilon\}) = \sum_{\varepsilon_j\in A} p_j$$

Eine endliche Verteilung

$$\mathrm{P}(\{\varepsilon_j\}) = 1/N\,, \qquad j = 1, \ldots, N$$

heißt **Gleichverteilung** oder **uniforme** Verteilung.

Lemma

In einem gleichverteilten Wahrscheinlichkeitsraum $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ gilt:

$$A \in \mathfrak{E} \quad \Rightarrow \quad \mathrm{P}(A) = \frac{n_A}{N} = \frac{|A|}{|\mathfrak{V}|} = \frac{\#(\text{„günstige Fälle“})}{\#(\text{„alle Fälle“})}$$

Prinzip vom unzureichenden Grunde

Principle of insufficient reason

Laplace-Prinzip

Haben wir keinen Anlaß, etwas anderes (Klügeres?) anzunehmen, so betrachten wir die Elementarereignisse eines W.-Raumes als gleichwahrscheinlich.

Klassische Wahrscheinlichkeitstheorie

Probabilistisches Schließen heißt **Zählen**!
Mathematisches Hilfsmittel ist die **Kombinatorik**.

Problem

Wie sieht eine unendliche Gleichverteilung aus?
Wie groß ist die Wahrscheinlichkeit, aus $\mathbb{N}$ eine gerade Zahl zu ziehen?

$$\mathbb{N} = \{1, 2, 3, 4, 6, 5, 8, 10, 12, 7, 14, 16, 18, 20, 9, 22, 24, 26, 28, 30, 11, 32, 34, 36, \ldots\}$$

Axiomatische Wahrscheinlichkeitstheorie

Bedingte Wahrscheinlichkeiten & Teilraumbildung

Definition

Sei $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathfrak{E}$ mit $\mathrm{P}(B) \neq 0$. Dann heißt

$$\mathrm{P}(A|B) \;=\; \frac{\mathrm{P}(A,B)}{\mathrm{P}(B)}$$

die **bedingte Wahrscheinlichkeit** des Ereignisses A unter der Bedingung B.

Satz

Sei $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ ein Wahrscheinlichkeitsraum und $B \in \mathfrak{E}$ mit $\mathrm{P}(B) \neq 0$. Dann ist auch $(\mathfrak{V}, \mathfrak{E}, \mathrm{P}^B)$ mit

$$\mathrm{P}^B(A) \;\stackrel{def}{=}\; \mathrm{P}(A|B) \qquad \textit{für alle } A \in \mathfrak{E}$$

ein Wahrscheinlichkeitsraum.

Axiomatische Wahrscheinlichkeitstheorie

Bayesformel & Kettenregel

Lemma (Bayesformel)

Für Ereignisse $A, B \in \mathfrak{E}$ mit $\mathrm{P}(A) \neq 0 \neq \mathrm{P}(B)$ gilt die Gleichung

$$\mathrm{P}(B|A) \;=\; \frac{\mathrm{P}(B,A)}{\mathrm{P}(A)} \;=\; \frac{\mathrm{P}(A,B)}{\mathrm{P}(A)} \;=\; \frac{\mathrm{P}(A|B)\cdot\mathrm{P}(B)}{\mathrm{P}(A)}$$

*Handelt es sich bei B um eine Ursache und bei A um eine Wirkung, so bezeichnen wir $\mathrm{P}(B|A)$ auch als **a posteriori** Wahrscheinlichkeit (von B unter A).*

Lemma (Kettenregel)

Für die Ereignisse $A_1, \ldots, A_n$ eines Wahrscheinlichkeitsraumes sei $\mathrm{P}(A_1 \ldots A_{n-1}) \neq 0$. Dann gilt die Faktorzerlegung

$$\begin{aligned}\mathrm{P}(A_1 \ldots A_n) \;=\;& \mathrm{P}(A_1)\cdot\mathrm{P}(A_2|A_1)\cdot\mathrm{P}(A_3|A_1A_2)\cdot\ldots \\ & \ldots\cdot\mathrm{P}(A_4|A_1A_2A_3)\cdot\ldots\cdot\mathrm{P}(A_n|A_1\ldots A_{n-1})\end{aligned}$$

Axiomatische Wahrscheinlichkeitstheorie

Marginalisierung und totale Wahrscheinlichkeit

Lemma

Im Wahrscheinlichkeitsraum $(\mathfrak{V}, \mathfrak{E}, \mathrm{P})$ gelte für $\mathfrak{V}$ die paarweise disjunkte Zerlegung $\mathfrak{V} = \bigcup_{i=1}^{n} B_i$. Dann gilt für jedes Ereignis A die Gleichung

$$\mathrm{P}(A) \;=\; \sum_{i=1}^{n}\mathrm{P}(A,B_i) \;=\; \sum_{i=1}^{n}\mathrm{P}(A|B_i)\cdot\mathrm{P}(B_i)$$

*der **totalen Wahrscheinlichkeit** sowie für jedes $j = 1, \ldots, n$ die Mehrwege-Bayesformel*

$$\mathrm{P}(B_j|A) \;=\; \frac{\mathrm{P}(A|B_j)\cdot\mathrm{P}(B_j)}{\sum_{i=1}^{n}\mathrm{P}(A|B_i)\cdot\mathrm{P}(B_i)}\;.$$

Axiomatische Wahrscheinlichkeitstheorie

Statistische und paarweise statistische Unabhängigkeit

Definition

Zwei Ereignisse A, B heißen **paarweise** (statistisch) **unabhängig** ($A \not\sim B$), wenn eine der folgenden Bedingungen erfüllt ist:

$$\begin{aligned}&(1) \quad \mathrm{P}(AB) \;=\; \mathrm{P}(A)\cdot\mathrm{P}(B)\\ &(2) \quad \mathrm{P}(A|B) \;=\; \mathrm{P}(A) \quad \text{und } \mathrm{P}(B)\neq 0\\ &(3) \quad \mathrm{P}(B|A) \;=\; \mathrm{P}(B) \quad \text{und } \mathrm{P}(A)\neq 0\end{aligned}$$

Die Ereignisse $A_1, \ldots, A_n \in \mathfrak{E}$ heißen statistisch **unabhängig** genau dann, wenn für alle denkbaren Indexkombinationen gilt:

$$\begin{aligned}\mathrm{P}(A_iA_j) &= \mathrm{P}(A_i)\cdot\mathrm{P}(A_j)\\ \mathrm{P}(A_iA_jA_k) &= \mathrm{P}(A_i)\cdot\mathrm{P}(A_j)\cdot\mathrm{P}(A_k)\\ \vdots &= \vdots\\ \mathrm{P}(\textstyle\bigcap A_\ell) &= \displaystyle\prod_{\ell}\mathrm{P}(A_\ell)\end{aligned}$$

Beispiele

- Paarweise statistische Unabhängigkeit (Würfelpaar-Versuch):

$$\begin{aligned} A &= \text{„die Augensumme ist ungerade“} \\ B &= \text{„der erste Wurf war eine '6'“} \\ C &= \text{„die Augensumme beträgt sieben“} \end{aligned}$$

  Dann gilt $A \not\sim B$, $B \not\sim C$, aber $A \sim C$.

- *Achtung* — aus der paarweisen Unabhängigkeit folgt im allgemeinen nicht die (allgemeine) Unabhängigkeit. Die Ereignisse

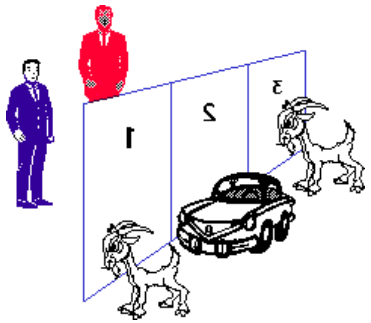
$$\begin{aligned} A_1 &= \text{„der erste Wurf war ungerade“} \\ A_2 &= \text{„der zweite Wurf war ungerade“} \\ A_3 &= \text{„die Summe war ungerade“} \end{aligned}$$

  sind zwar paarweise unabhängig, doch es gilt

$$\mathrm{P}(A_1 A_2 A_3) \;=\; 0 \;\neq\; {}^1\!/_8 \;=\; \mathrm{P}(A_1) \cdot \mathrm{P}(A_2) \cdot \mathrm{P}(A_3)$$

Das Monty-Hall-Paradoxon

Der Fachmann staunt ... und der Laie wundert sich!



Fragestellung

Der Moderator einer Spielshow zeigt dem Kandidaten 3 Türen. „Hinter einer der drei Türen steht der Hauptgewinn, ein Volvo. Hinter den anderen beiden Türen stehen Ziegen. Welche Tür wählen Sie?“ Nachdem der Kandidat sich entschieden hat, öffnet der Moderator eine der beiden anderen Türen — mit einer stinkenden Ziege dahinter! „Bleiben Sie bei Ihrer Wahl oder möchten Sie noch einmal umwählen?“

Zusammenfassung (6)

1. Ziel des Klassifikatorentwurfs ist die **Minimierung** der zu erwartenden **Fehlerrate**.
2. Der Klassifikator wird mit Methoden des **maschinellen Lernens** aus einer etikettierten **Lernstichprobe** gewonnen.
3. Laut statistischer **Entscheidungstheorie** wird die minimale Fehlerrate durch die **Bayesregel** (MAP-Regel) garantiert.
4. Die Bayesregel ist **nicht praktikabel**, denn sie erfordet die exakte Kenntnis des **Mustererzeugungsprozesses**.
5. **Statistische Klassifikatoren** wie der uniform-naive Bayesklassifikator schätzen ein Näherungsmodell der Musterverteilung.
6. **Diskriminative Klassifikatoren** wie die lineare Quadratmittel-Trennfunktion approximieren die a posteriori Klassenwahrscheinlichkeiten.
7. **Nichtparametrische Klassifikatoren** wie die kNN-Regel verzichten auf explizite Strukturannahmen; der Klassifikationsaufwand steigt dann mit dem Lernprobenumfang.
8. Eine faire Schätzung der **Klassifikatorleistung** erfordert strikte **Trennung von Lern- und Testdaten**, z.B. mittels Kreuzvalidierung (L^1O).