MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2017

Prof. E.G. Schukat-Talamazzini

Stand: 13. Oktober 2017

Teil IV

Vorhersage und Kategorisierung

Vorhersage und statistische Abhängigkeit

Charakterisierung der statistischen Unabhängigkeit

Zwei Variablenmengen $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_N)$ und $\mathbb{Y} = (\mathbb{Y}_1, \ldots, \mathbb{Y}_M)$ heißen **statistisch unabhängig** voneinander gdw. gilt:

$$(\forall \boldsymbol{x})(\forall \boldsymbol{y}) \qquad \mathrm{P}(\mathbb{X} = \boldsymbol{x}, \mathbb{Y} = \boldsymbol{y}) \;=\; \mathrm{P}(\mathbb{X} = \boldsymbol{x}) \cdot \mathrm{P}(\mathbb{Y} = \boldsymbol{y})$$

Für Tupel $\boldsymbol{x}$ mit $\mathrm{P}(\boldsymbol{x}) \neq 0$ ist das äquivalent zu:

$$\mathrm{P}(\mathbb{Y} = \boldsymbol{y} \mid \mathbb{X} = \boldsymbol{x}) \;=\; \mathrm{P}(\mathbb{Y} = \boldsymbol{y})$$

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \;\Rightarrow\; \boxed{\begin{array}{c}\text{Datenmodell}\\ \mathrm{P}(\mathbb{Y} = \boldsymbol{y} \mid \mathbb{X} = \boldsymbol{x})\end{array}} \;\Rightarrow\; (\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_M)$$

Fakt

*Im Fall statistischer Abhängigkeit besteht eine Chance, die Werte der **endogenen** Variablen $\mathbb{Y}_m$ aus den Werten der **exogenen** Variablen $\mathbb{X}_n$ zu „erraten“.*

Statistische Prädiktion von Einzelvariablen

Quellvariable $(\mathbb{X}_1, \ldots, \mathbb{X}_N)$ ➪ Zielvariable $\mathbb{Y}_1 =: \mathbb{Y}$

Maschinelles Lernen eines Vorhersagemodells

Entscheidungsfunktion: $f : \mathcal{X}_1 \times \ldots \times \mathcal{X}_N \to \mathcal{Y}$

Kostenfunktion („*loss*"): $\mathcal{L}(\boldsymbol{x}, y, \hat{y})$ mit $\hat{y} = f(\boldsymbol{x})$

Risiko (zu minimieren): $\mathfrak{R}(f) := \mathcal{E}_{\mathrm{P}(\boldsymbol{x},y)}[\mathcal{L}(\mathbb{X}, \mathbb{Y}, f(\mathbb{Y})]$

$\mathbb{Y}$ nominal

$\mathcal{L}(\boldsymbol{x}, y, \hat{y}) = c_{y\hat{y}}$

Kostenmatrix $\boldsymbol{C}$ mit $c_{\kappa\kappa} \leq c_{\kappa\lambda}$

$\mathbb{Y}$ ordinal

$\mathcal{L}(\boldsymbol{x}, y, \hat{y}) = c_{y\hat{y}}$

Diskrepanzmatrix $\boldsymbol{C}$ mit $c_{k\ell} \leq c_{k'\ell'}$ für $k' \leq k \leq \ell \leq \ell'$

$\mathbb{Y}$ kardinal

$\mathcal{L}(\boldsymbol{x}, y, \hat{y}) = d(y, \hat{y})$

metrische Distanzmaße $d(y, \hat{y}) = |y - \hat{y}|^p$, $p \geq 0$

Spezialfall

(Fehlerrate)

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

Spezialfall

(Linearskala)

$c_{k\ell} = |z_k - z_\ell|$

Spezialfall

(Quadratmittel)

$d(y, \hat{y}) = (y - \hat{y})^2$

Optimale Prädiktion in den Spezialkonfigurationen

$$\mathfrak{R}(f) = \mathcal{E}[\mathcal{L}(\mathbb{X}, \mathbb{Y}, f(\mathbb{Y})] = \int \sum_y \mathrm{P}(\boldsymbol{x}, y) \cdot c_{y, f(x)} \, dx$$

Klassifikation (Bayesregel)

$\mathbb{Y}$ ist **nominal** *Modus*

$$\hat{y}(\boldsymbol{x}) = \operatorname*{argmax}_{\kappa \in \boldsymbol{\Omega}_y} \mathrm{P}(\mathbb{Y} = \kappa \mid \boldsymbol{x})$$

Ordinale Klassifikation

$\mathbb{Y}$ ist **ordinal** *Median*

$$\hat{y}(\boldsymbol{x}) = \operatorname*{median}_{\ell \in \boldsymbol{\Omega}_y} \mathrm{P}(\mathbb{Y} = \ell \mid \boldsymbol{x})$$

Quadratmittel-Regression

$\mathbb{Y}$ ist **kardinal** *Mean*

$$\hat{y}(\boldsymbol{x}) = \mathcal{E}_{\mathbb{Y}|\boldsymbol{x}}[\mathbb{Y}] = \int_{\mathrm{IR}} \mathrm{P}(y|\boldsymbol{x}) \cdot y \, dy$$

Klassifikationsverfahren

Welche(n) Skalentyp(en) besitzen die exogenen Variablen ?

Numerisch

NV-Klassifikator
Polynomklassifikator
Multilayer-Perzeptron
Supportvektormaschine

Diskret

Versionenraumverfahren
Kanonische+naive Bayesregel
Markovnetze

Metrisch

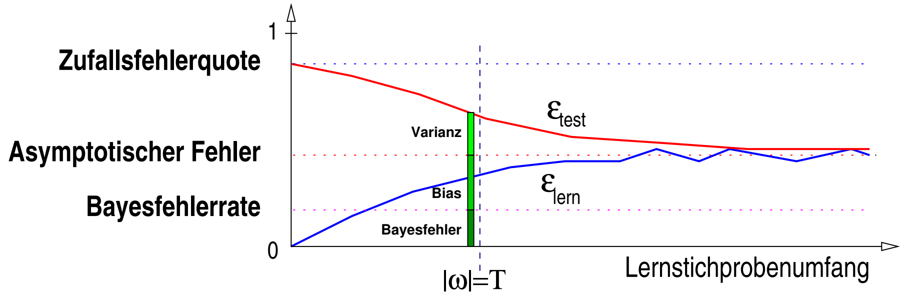
Nächste-Nachbar-Regeln
SVM + Kerneltrick
MDS + **numerisch**

Numerisch & diskret

Entscheidungsbäume
Loglinearmodelle
Bayesnetze
(Konversion)

Fehlerrate, Überanpassung & Unteranpassung

Was wir schon in der Vorlesung „*Mustererkennung*" über das Lernen gelernt haben



Lernstichprobe des Klassifikationsverfahrens $\omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$
Fehlerrate auf den Lerndaten $\varepsilon_{\text{lern}}$
Fehlerrate auf den Testdaten (≈ Fehlerwahrscheinlichkeit) $\varepsilon_{\text{test}}$

Bayesfehler — weniger geht nicht
Zufallsfehler — mehr muss nicht
Grenzfehler — Daten! Daten!!

Bias

Datenmodell ⇩

Varianz

Lernprobe ⇩

Begriffe (Konzepte)

Intensionaler Zugehörigkeitstest: $\phi(\boldsymbol{x}) = \phi(x_1, \ldots, x_N) = 1$?

Scharf oder unscharf?

Ist dieses Element aus

$$\boldsymbol{\Omega} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N$$

ein $\left\{ \begin{array}{l} \text{Tiger ?} \\ \text{Liger ?} \\ \text{Töwe ?} \\ \text{Löwe ?} \end{array} \right\}$

Extension eines Begriffs

Mengentheoret. charakterisiert

$$\mathcal{C} \subseteq \boldsymbol{\Omega}$$

$\boldsymbol{x} \in \mathcal{C} \Leftrightarrow \boldsymbol{x}$ *„ist“* ein $\mathcal{C}$

Intension eines Begriffs

Algorithmisch charakterisiert

$$\phi : \boldsymbol{\Omega} \to \{0, 1\}$$

„Parser“ ϕ entscheidbare Funkt.
($\rightsquigarrow$ endlich aufschreibbar)

Kardinalitätskonflikt

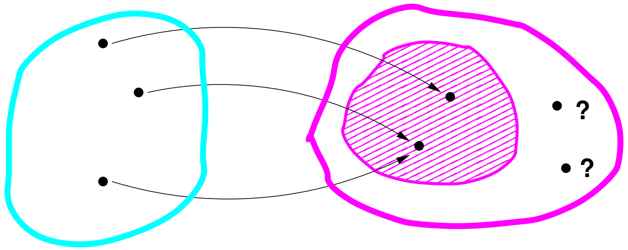
„Worüber man nicht reden kann, darüber soll man schweigen.“

Intension

$\{\phi \mid \phi : \boldsymbol{\Omega} \to \{0, 1\}$ terminierender Algorithmus$\}$

Extension

Potenzmenge $\mathfrak{P}\boldsymbol{\Omega}$



abzählbar unendlich

$\langle 0 \rangle, \langle 1 \rangle,$
$\langle 00 \rangle, \langle 01 \rangle, \langle 10 \rangle, \langle 11 \rangle,$
$\langle 000 \rangle, \langle 001 \rangle, \langle 010 \rangle, \ldots,$
$\langle 0000 \rangle, \langle 0001 \rangle, \langle 0010 \rangle, \ldots,$
$\langle 00000 \rangle, \ldots, \langle 000000 \rangle, \ldots$

überabzählbar unendlich

$\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \ldots$
$\{x_1, x_2\}, \{x_1, x_3\}, \ldots, \{x_2, x_3\}, \ldots$
$\{x_1, x_2, x_3\}, \{x_1, x_2, x_4\}, \ldots$
$\{x_1, x_2, x_3, x_4\}, \{x_1, x_2, x_3, x_5\}, \ldots$
$\ldots \ldots \ldots$

$$\phi \quad \mapsto \quad \boldsymbol{\Omega}_\phi \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \boldsymbol{\Omega} \mid \phi(\boldsymbol{x}) = 1\} \in \mathfrak{P}\boldsymbol{\Omega}$$

Abstraktion

Konzentration auf die Objekteigenschaften zu Lasten der Objektidentität

Gegeben

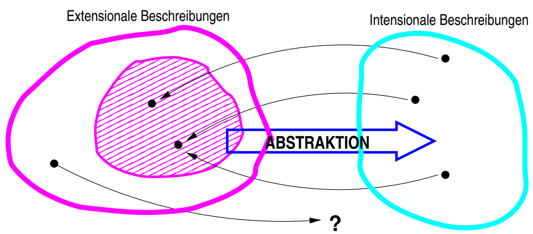
ist ein Konzept
$\mathcal{C} \subseteq \boldsymbol{\Omega}$ (extensional)

Gesucht

ist eine *„krispe“* intensionale Beschreibung
$\phi : \boldsymbol{\Omega} \to \{0, 1\}$ mit der Eigenschaft

$$\boldsymbol{x} \in \mathcal{C} \Leftrightarrow \phi(\boldsymbol{x}) = 1$$

($\rightsquigarrow$ Kompatibilität: $\boldsymbol{\Omega}_\phi = \mathcal{C}$)



Problem

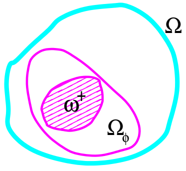
Nicht jedes Konzept ist abstrahierbar!

Induktion

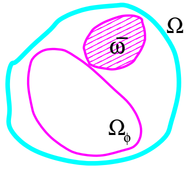
Verallgemeinerung oder „*Lernen aus Beispielen*“

Gegeben

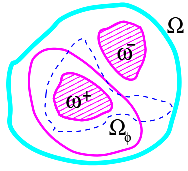
Positivbeispiele
$\omega^+ \subseteq \boldsymbol{\Omega}_\phi$
Negativbeispiele
$\omega^- \subseteq \boldsymbol{\Omega} \setminus \boldsymbol{\Omega}_\phi$



positive Beispiele



negative Beispiele



positive und negative Beispiele

Gesucht

kompatible intensionale Beschreibung ψ:

$$(\forall \boldsymbol{x} \in \boldsymbol{\Omega}) \quad \boldsymbol{x} \in \omega^+ \quad \Rightarrow \quad \psi(\boldsymbol{x}) = 1$$
$$(\forall \boldsymbol{x} \in \boldsymbol{\Omega}) \quad \boldsymbol{x} \in \omega^- \quad \Rightarrow \quad \psi(\boldsymbol{x}) = 0$$

Potentielle Kandidaten sind alle ψ mit

$$\omega^+ \subseteq \boldsymbol{\Omega}_\psi \subseteq \boldsymbol{\Omega} \setminus \omega^-$$

Lernverfahren

- Hypothesenraum
- Lösungsvielfalt
- Auswahlkriterium

Induktionsproblematik

Lerndaten? Hypothesen? Verfahren?

Übergeneralisierung

Es werden *Oberbegriffe* von ϕ gelernt.

Abhilfe

Negativbeispiele bereitstellen

Überspezialisierung

Es werden *Unterbegriffe* von ϕ gelernt.

Abhilfe

Repräsentative Positivstichprobe

Fehlgranulation

Überanpassung oder *Unteranpassung*

Abhilfe

Adäquates Sortiment von Hypothesen

Natürlichkeit, Fortsetzbarkeit

Gelernte Verallgemeinerung versagt bei Wiederabruf

Abhilfe

Occam's razor: einfache Erklärung

Kategorisierung von Objekten

(a.k.a. „*Klassifikation*“)

Extensionale Charakt. *vs.*

Mengenpartition
$\mathcal{C}_1, \ldots, \mathcal{C}_K \subseteq \boldsymbol{\Omega}$ mit

$$\bigcup_{\kappa=1}^{K} \mathcal{C}_\kappa = \boldsymbol{\Omega}$$

und für alle $\kappa \neq \lambda$:

$$\mathcal{C}_\kappa \cap \mathcal{C}_\lambda = \varnothing$$

Spezialfall $K = 2$

Konzept $\mathcal{C}_1 = \mathcal{C}$ und sein
Komplement $\mathcal{C}_2 = \boldsymbol{\Omega} \setminus \mathcal{C}$

Intensionale Charakterisierung

keine wirklich zwingende Verallgemeinerung:

1. **Charakteristische Fkt.**
   $\phi : \boldsymbol{\Omega} \to \{1, \ldots, K\} \subset \mathbb{R}$
2. **Konzepttupel**
   $\boldsymbol{\phi} : \boldsymbol{\Omega} \to \{0,1\}^K$ mit $\boldsymbol{\Omega}_{\phi_\kappa} \stackrel{!}{=} \mathcal{C}_\kappa$
3. **Diskriminanten**
   $\boldsymbol{\phi} : \boldsymbol{\Omega} \to \mathbb{R}^K$ mit $\boldsymbol{x} \in \mathcal{C}_\kappa$ gdw.
   $$\phi_\kappa(\boldsymbol{x}) \geq \phi_\lambda(\boldsymbol{x}) \quad (\forall \lambda)$$
4. **Nominale Regression**
   $\mathrm{P} : \boldsymbol{\Omega} \times \{\xi_1, \ldots, \xi_K\} \to \mathbb{R}$
   plus Bayesregel

Aussagenlogisches Lernen

Begriffe lernen · Klassifikation · Gruppierung

Segel-Szenarium

$\mathcal{X}_1$ sky	$\mathcal{X}_2$ air	$\mathcal{X}_3$ humidity	$\mathcal{X}_4$ wind	$\mathcal{X}_5$ water	$\mathcal{X}_6$ forecast
{sunny, rainy, cloudy}	{warm, cool}	{normal, high, low}	{strong, weak}	{warm, cool}	{same, change}

Single Representation Trick

z.B. Hypothesen als unvollständige Attributwertspezifikationen

- **Objekte** ≙ Attributbelegungen

$$(\mathit{sunny}, \mathit{warm}, \mathit{normal}, \mathit{strong}, \mathit{warm}, \mathit{same}) \in \boldsymbol{\Omega}$$

- **Hypothesen** ≙ partielle Attributbelegungen

$$(\mathit{sunny}, ?, ?, \mathit{strong}, ?, ?) \in \mathcal{H}$$

Hypothesenraum

Konjunktionen positiver Literale (KPL)

Definition

Es sei $\boldsymbol{\Omega} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_N$ ein Objektraum. Dann heißen die Elemente aus

$$\mathcal{H} = (\mathcal{X}_1 \cup \{?\}) \times \ldots \times (\mathcal{X}_N \cup \{?\})$$

KPL-Hypothesen über $\boldsymbol{\Omega}$. Die Menge $\mathcal{H}$ heißt **KPL-Hypothesenraum** über $\boldsymbol{\Omega}$.

Ein Beispielobjekt $\boldsymbol{x} \in \boldsymbol{\Omega}$ genügt der Hypothese $h \in \mathcal{H}$ ($\boldsymbol{x}$ **erfüllt** h bzw. $h \models \boldsymbol{x}$) genau dann, wenn gilt:

$$\forall i = 1, \ldots, N : \quad (h_i = ?) \vee (h_i = x_i)$$

Bemerkung

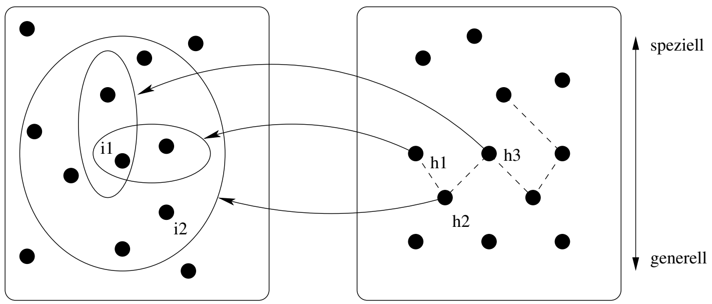
Vollständige KPL ≙ Objekte
Leere KPL ≙ Konzept $\mathcal{C} = \boldsymbol{\Omega}$
Definiere $h_\varnothing$ ≙ Leerkonzept $\mathcal{C} = \varnothing$

Segel-Szenarium

Objektraum: $|\boldsymbol{\Omega}| = 144$
KPL-Hypothesenraum: $|\mathcal{H}| = 1296$
Konzeptraum:
$|\mathfrak{P}\boldsymbol{\Omega}| = 2^{144} \approx 1000^{14.4} \approx 10^{43}$

Hypothesen und Objektmengen

$\boldsymbol{\Omega}(h) \mathrel{\hat=} \{\boldsymbol{x} \in \boldsymbol{\Omega} \mid h \models \boldsymbol{x}\}$



Segel-Szenarium

i_1 : $(\mathit{sunny}, \mathit{warm}, \mathit{high}, \mathit{strong}, \mathit{cool}, \mathit{same})$
i_2 : $(\mathit{sunny}, \mathit{warm}, \mathit{high}, \mathit{light}, \mathit{warm}, \mathit{same})$

h_1 : $(\mathit{sunny}, ?, ?, \mathit{strong}, ?, ?)$
h_2 : $(\mathit{sunny}, ?, ?, ?, ?, ?)$
h_3 : $(\mathit{sunny}, ?, ?, ?, \mathit{cool}, ?)$

$h_2 \supseteq h_1, h_3$
bzw.
$h_1, h_3 \Rightarrow h_2$

Der Verband aller KPL-Hypothesen

Definition

Für jede Hypothese $h \in \mathcal{H}$ sei $\boldsymbol{\Omega}(h) \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \boldsymbol{\Omega} \mid h \models \boldsymbol{x}\}$ (Extension) definiert. Die Menge $\mathcal{H}$ erbt von $\mathfrak{P}\boldsymbol{\Omega}$ die **Inklusionsrelation** (h ist „*allgemeiner*“ oder „*genereller*“ als h'):

$$h \supseteq h' \quad \Leftrightarrow \quad \forall \boldsymbol{x} \in \boldsymbol{\Omega} : \; (h' \models \boldsymbol{x} \Rightarrow h \models \boldsymbol{x})$$

Der Raum aller DNF-Hypothesen (disjunktive Normalform) ist die Boolesche Algebra $(\mathfrak{P}\boldsymbol{\Omega}, \subseteq)$.

Lemma

Der Raum $(\mathcal{H}, \subseteq)$ bildet eine Halbordnung. {*reflexiv*, *transitiv*, *antisymmetrisch*}

Die KPL-Hypothesen sind abgeschlossen gegenüber Durchschnittbildung.

Die KPL-Hypothesen sind nicht abgeschlossen gegenüber der Mengenvereinigung, es existiert das Supremum je zweier Hypothesen:

$$(h \vee h')_n \stackrel{\text{def}}{=} \begin{cases} v & (\exists v \in \mathcal{X}_n)\; h_n = v = h'_n \\ ? & h_n \neq h'_n \\ ? & h_n = ? = h'_n \end{cases}$$

Sukzessiver Generalisierungsalgorithmus

Definition

Eine Hypothese $h \in \mathcal{H}$ heißt **konsistent** mit den Lerndaten (ω^+, ω^-) genau dann wenn gilt:

$$x \in \omega^+ \quad \Rightarrow \quad h \models x$$
$$x \in \omega^- \quad \Rightarrow \quad h \not\models x$$

Bemerkungen

1. Konsistenz falls $\omega^+ \subseteq \boldsymbol{\Omega}_h \subseteq \boldsymbol{\Omega} \setminus \omega^-$
2. Jedes h ist konsistent mit $(\varnothing, \varnothing)$.
3. Kein h ist konsistent wenn $\omega^+ \cap \omega^- \neq \varnothing$.
4. Auch für disjunkte (ω^+, ω^-) enthält $\mathcal{H}$ nicht notwendig eine konsistente Hypothese!

(1) INITIALISIERUNG
Setze $h \leftarrow h_\varnothing$.

Keine $\boldsymbol{x} \in \omega^-$ verwendet.

(2) GENERALISIERUNG
Setze für alle $\boldsymbol{x} \in \omega^+$:

$$h \leftarrow h \vee \boldsymbol{x}$$

(*„speziellste Erweiterung“* von h um $\boldsymbol{x}$)

Resultat genügt allen $\boldsymbol{x} \in \omega^+$. h ist minimal mit dieser Eigenschaft.

Wenn konsistente Hypothese existiert, wird sie gefunden.

(3) TERMINIERUNG
Das Ergebnis ist h.

Nur für KPL (Supremum!) realisierbar.
Für $\mathcal{H} = \mathfrak{P}\boldsymbol{\Omega}$ ist SGA trivial.

Kandidateneliminationsalgorithmus

Suche operiert auf (*„Kandidaten“*-) Mengen von Hypothesen

(1) INITIALISIERUNG
Setze $H \leftarrow \mathcal{H}$

(2) GENERALISIERUNG / SPEZIALISIERUNG
Eliminiere für alle $\boldsymbol{x} \in \omega^+ \cup \omega^-$

(a) Fall $\boldsymbol{x} \in \omega^+$: alle $h \in H$ mit $h \not\models \boldsymbol{x}$
(b) Fall $\boldsymbol{x} \in \omega^-$: alle $h \in H$ mit $h \models \boldsymbol{x}$

(3) TERMINIERUNG
Das Ergebnis ist h, falls $H = \{h\}$ ist.

Am Ende enthält die Kandidatenmenge genau die konsistenten Hypothesen aus $\mathcal{H}$.

Es gibt keine, eine oder mehrere Lösungen.

Das Verfahren ist aus Aufwandsgründen impraktikabel!

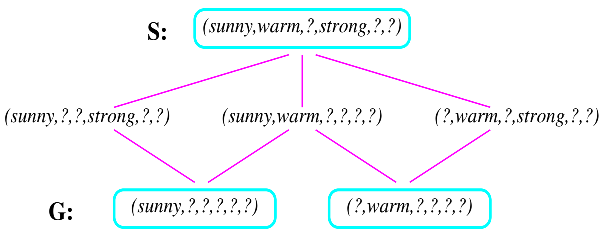
Der Versionenraum

Definition

Die Menge der mit den Lernbeispielen konsistenten Hypothesen

$$\{h \in \mathcal{H} \mid h \text{ konsistent mit } (\omega^+, \omega^-)\}$$

heißt **Versionenraum** von (ω^+, ω^-) bezüglich $\mathcal{H}$ und wird mit $\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-)$ (oder $\mathfrak{V}$) bezeichnet.



Beispiel

Versionenraum mit 6 Hypothesen

1x minimal
2x maximal
3x weder/noch

Minimale und maximale VR-Elemente

$$\mathfrak{V}_S \stackrel{\text{def}}{=} \{h \in \mathfrak{V} \mid \forall h' \in \mathfrak{V}: \ h' \subseteq h \Rightarrow h' = h\}$$
$$\mathfrak{V}_G \stackrel{\text{def}}{=} \{h \in \mathfrak{V} \mid \forall h' \in \mathfrak{V}: \ h \subseteq h' \Rightarrow h' = h\}$$

Versionenräume als Halbordnungsintervalle

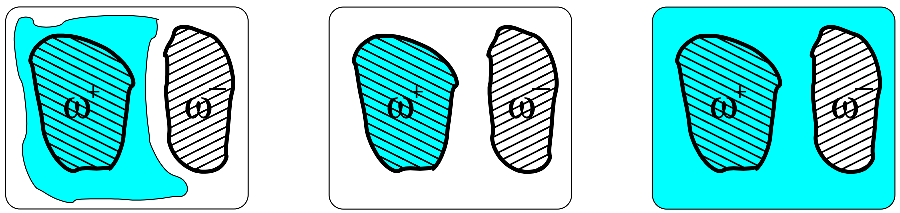
Beispiel

Im vollständigen Hypothesenraum $\mathcal{H} = \mathfrak{P}\boldsymbol{\Omega}$ sind die minimalen und die maximalen VR-Elemente eindeutig:

$$\mathfrak{V}_S = \{\omega^+\} \qquad \text{und} \qquad \mathfrak{V}_G = \{\boldsymbol{\Omega} \setminus \omega^-\}$$

Versionenräume besitzen die Gestalt einer **Intervalldarstellung**:

$$\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-) = \{h \in \mathcal{H} \mid \omega^+ \subseteq h \subseteq \boldsymbol{\Omega} \setminus \omega^-\} = [\mathfrak{V}_S, \mathfrak{V}_G]_{\mathcal{H}}$$



eine VR-Hypothese — die kleinste VR-Hypothese — die größte VR-Hypothese

Der Versionenraum-Darstellungssatz

Die Intervalldarstellung gilt in allen beliebigen Hypothesenräumen

Definition

Sei $\mathcal{H} \subseteq \mathfrak{P}\mathbf{\Omega}$; ein **einfaches HO-Intervall** in $\mathcal{H}$ hat die Form:

$$[h_u, h_o]_\mathcal{H} \stackrel{\text{def}}{=} \{h \in \mathcal{H} \mid h_u \subseteq h \subseteq h_o\}$$

Ein **verallgemeinertes HO-Intervall** in $\mathcal{H}$ hat die Form:

$$[\mathcal{H}_u, \mathcal{H}_o]_\mathcal{H} = \{h \in \mathcal{H} \mid \exists h_u \in \mathcal{H}_u, \exists h_o \in \mathcal{H}_o : h_u \subseteq h \subseteq h_o\}$$

$$[\mathcal{H}_u, \mathcal{H}_o]_\mathcal{H} = \bigcup_{h_u \in \mathcal{H}_u} \bigcup_{h_o \in \mathcal{H}_o} [h_u, h_o]_\mathcal{H}$$

Satz

Für den Versionenraum $\mathfrak{V}$ der Beispieldaten ω^+ und ω^- bezüglich $\mathcal{H}$ gilt eine Intervalldarstellung:

$$\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-) = [\mathfrak{V}_S, \mathfrak{V}_G]_\mathcal{H}$$

Dabei sind $\mathfrak{V}_S$ und $\mathfrak{V}_G$ die Mengen der $\subseteq$-minimalen ($\subseteq$-maximalen) Elemente des Versionenraums $\mathfrak{V}$.

Beweis.

Inklusionsrichtung $\subseteq$:

Sei $h \in \mathfrak{V}$. Sei $G(h) := \{h' \in \mathfrak{V} | h' \supseteq h\}$.
Wegen $h \in G(h)$ ist $G(h) \neq \varnothing$.

- Sei h_G ein maximales Element aus $G(h)$.
  Dann ist $h_G \in \mathfrak{V}_G$ und $h_G \supseteq h$.

(Die Existenz eines $h_S \in \mathfrak{V}_S$ zeigt man/frau analog.)

Inklusionsrichtung $\supseteq$:

Sei $h \in \mathcal{H}$ mit $h \subseteq h_G \in \mathfrak{V}_G$ und $h \supseteq h_S \in \mathfrak{V}_S$.
Zu zeigen: h ist konsistent mit (ω^+, ω^-), d.h. $h \in \mathfrak{V}$.

1. Sei $x \in \omega^+$.
   Wegen $h_S \in \mathfrak{V}_S \subseteq \mathfrak{V}$ gilt $h_S \models x$.
   Wegen $h \supseteq h_S$ gilt auch $h \models x$.
2. Sei $x \in \omega^-$.
   Wegen $h_G \in \mathfrak{V}_G \subseteq \mathfrak{V}$ gilt $h_G \not\models x$.
   Wegen $h \subseteq h_G$ gilt auch $h \not\models x$.

□

Versionenraum-Kandidateneliminationsalgorithmus

(Algorithmus)

1 INITIALISIERUNG
Setze $G \leftarrow \{\mathbf{\Omega}\}$ und $S \leftarrow \{\varnothing\}$.

2^+ POSITIVE BEISPIELE
Für alle $\boldsymbol{x} \in \omega^+$:
· Entferne alle $h \in G$ mit $h \not\models \boldsymbol{x}$
· Für alle $h \in S$:
  Generalisiere h zu h' mit $h' \models \boldsymbol{x}$
  Behalte $h' \in S$, falls h' spezieller als G
· Entferne alle nichtminimalen $h \in S$

2^- NEGATIVE BEISPIELE
Für alle $\boldsymbol{x} \in \omega^-$:
· Entferne alle $h \in S$ mit $h \models \boldsymbol{x}$ · Für alle $h \in G$:
  Spezialisiere h zu h' mit $h' \not\models \boldsymbol{x}$
  Behalte $h' \in G$, falls h' allgemeiner als S
· Entferne alle nichtmaximalen $h \in G$

3 TERMINIERUNG
Das Ergebnis ist h, falls $G = \{h\} = S$ ist.

Bemerkungen

1. Grundidee: alle Versionenräume werden als „Intervalle" $[S, G]$ abgespeichert, und auch die Hypothesenelimination geschieht auf S, G und nicht auf $\mathfrak{V}$.
2. Es gilt natürlich $\mathcal{H} = [\varnothing, \mathbf{\Omega}]_\mathcal{H}$.
3. Wenn es geeignete Hypothesen mit $\mathbf{\Omega}(h_\varnothing) = \varnothing$ und $\mathbf{\Omega}(h_\mathbf{\Omega}) = \mathbf{\Omega}$ gibt, kann entsprechend initialisiert werden.
4. Hypothesen $h \in G$, die einem Positivbeispiel $\boldsymbol{x} \in \omega^+$ nicht genügen, dürfen ohne weiteres eliminiert werden, da jegliche Spezialisierung von h ebenfalls an $\boldsymbol{x}$ scheitern würde. Dasselbe gilt für $h \in S$, $\boldsymbol{x} \in \omega^-$ mit $h \models \boldsymbol{x}$.
5. Gilt jedoch für $\boldsymbol{x} \in \omega^+$ und ein $h \in S$ die Aussage $h \not\models \boldsymbol{x}$, so darf h wegen der Gefährdung des Teilraums $[h, G]$ nicht einfach gelöscht werden!
6. Von allen Generalisierungen h' von h mit $h' \models \boldsymbol{x}$ interessieren natürlich nur diejenigen mit $[h', G] \neq \varnothing$, und die auch minimal sind in S mit dieser Eigenschaft.
7. Am Ende sind alle Hypothesen aus S und aus G und auch aus $[S, G]$ konsistent mit den Beispieldaten, und $[S, G]$ ist auch diesbezüglich vollständig.

Beispiel (Segeln im KPL-Hypothesenraum)

Versionenraum nach VRE-Algorithmus

	sky	*air*	*humidity*	*wind*	*water*	*forecast*
$h_1 \in S$	*sunny*	*warm*	?	*strong*	?	?
h_2	*sunny*	?	?	*strong*	?	?
h_3	*sunny*	*warm*	?	?	?	?
h_4	?	*warm*	?	*strong*	?	?
$h_5 \in G$	*sunny*	?	?	?	?	?
$h_6 \in G$	?	*warm*	?	?	?	?

Unbeobachtete („*neue*“) Objekte

Vorhersage des Konzepts „*go_sailing*“:

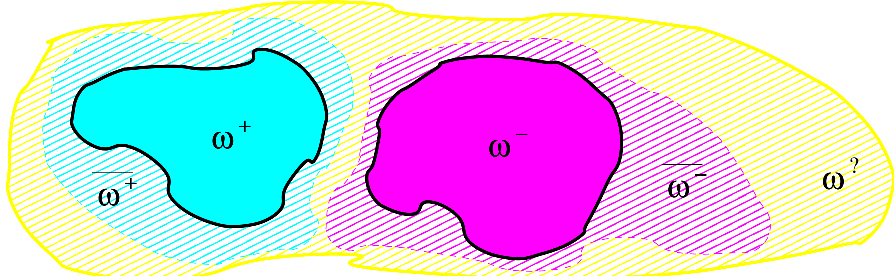
	sky	*air*	*hum*	*wind*	*water*	*fore*	S h_1	h_2	h_3	h_4	G h_5	G h_6
x_1	*sunny*	*warm*	*norm*	*strong*	*cool*	*change*	1	1	1	1	1	1
x_2	*rainy*	*cold*	*norm*	*weak*	*warm*	*same*	0	0	0	0	0	0
x_3	*sunny*	*warm*	*norm*	*weak*	*warm*	*same*	0	0	1	0	1	1

Parlamentarischer Alltag im Versionenraum

Positiver Konsens
Für alle $h \in \mathfrak{V}$ gilt
$h \models x$

Negativer Konsens
Für alle $h \in \mathfrak{V}$ gilt
$h \not\models x$

Ambiges Votum
Ex. $h_+, h_- \in \mathfrak{V}$ mit
$h_+ \models x$ und $h_- \not\models x$



Bemerkungen

1. Alle $h \in \mathfrak{V}$ sind konsistent ➪ $\left\{\begin{array}{l}\text{alle } \boldsymbol{x} \in \omega^+ \text{ werden einstimmig akzeptiert}\\ \text{alle } \boldsymbol{x} \in \omega^- \text{ werden einstimmig abgewiesen}\end{array}\right\}$.
2. Für $\mathfrak{V} = \varnothing$ folgt $\overline{\omega^+} \cap \overline{\omega^-} = \varnothing$.
3. Ist $\boldsymbol{x} \in \boldsymbol{\Omega}$ ambig, so ex. Hypothesen $h^+ \in \mathfrak{V}_G$, $h^- \in \mathfrak{V}_S$ mit $\left\{\begin{array}{l} h^+ \models \boldsymbol{x} \\ h^- \not\models \boldsymbol{x}\end{array}\right\}$.

Die induktive Hülle

Definition

Wir bezeichnen die Menge

$$\overline{\omega^+} \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \boldsymbol{\Omega} \mid h \in \mathfrak{V}(\omega^+, \omega^-) \Rightarrow h \models \boldsymbol{x}\}$$

als **induktive Hülle** der Positivbeispiele ω^+ und die Menge

$$\overline{\omega^-} \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \boldsymbol{\Omega} \mid h \in \mathfrak{V}(\omega^+, \omega^-) \Rightarrow h \not\models \boldsymbol{x}\}$$

als **induktive Hülle** der Negativbeispiele ω^-. Die Elemente aus

$$\omega^? \stackrel{\text{def}}{=} \boldsymbol{\Omega} \setminus \left(\overline{\omega^+} \cup \overline{\omega^-}\right)$$

heißen **ambige Objekte** von $\boldsymbol{\Omega}$ bezüglich $\mathcal{H}$, ω^+ und ω^-.

Lemma

Die Operatoren $\omega^+ \mapsto \overline{\omega^+}$ und $\omega^- \mapsto \overline{\omega^-}$ sind tatsächlich Hüllenoperatoren:

1. $\omega_1^+ \subseteq \omega_2^+$ ➪ $\overline{\omega_1^+} \subseteq \overline{\omega_2^+}$ *(Monotonie)*
2. $\omega^+ \subseteq \overline{\omega^+}$ *und* $\omega^- \subseteq \overline{\omega^-}$ *(Inklusion)*
3. $\overline{\omega^+} = \overline{\overline{\omega^+}}$ *und* $\overline{\omega^-} = \overline{\overline{\omega^-}}$ *(Involution)*

Der induktive Bias

Großartiger Lernerfolg durch mangelhafte Ausdrucksfähigkeit

Induktives Schließen

steht und fällt mit dem **Ausdrucksdefizit** des Hypothesenraums $\mathcal{H}$.

ω^+	*sunny*	*warm*	*normal*	*strong*	*warm*	*same*	
ω^+	*rainy*	*warm*	*normal*	*weak*	*warm*	*same*	
$\mathfrak{V}_S$	?	*warm*	*normal*	?	*warm*	*same*	
$\omega^?$	*sunny*	*warm*	*normal*	*weak*	*warm*	*same*	➪ +

Aussagenlogisch orientierte Hypothesenräume

- Konjunktion positiver Literale — KPL: $x_i = \xi$ oder ?
- Konjunktion positiver und negativer Literale — $x_i = \xi$ oder $x_i \neq \xi$ oder ?
- Konjunktion disjunktiver Komplexe — $x_i \in \mathcal{X}^+$
- Disjunktion positiver (und negativer) Literale — (dual zu oben)
- Disjunktion von Konjunktionen positiver Literale — $\mathcal{H} = \mathfrak{P}\boldsymbol{\Omega}$

Lernen einelementiger Versionenräume

Zur Auswahl neuer Lernbeispiele

- Erweiterung von ω^+ um Beispiele aus $\overline{\omega^+}$ ist überflüssig.
  Erweiterung von ω^- um Beispiele aus $\overline{\omega^-}$ ist überflüssig.
- Erweiterung von ω^+ um Beispiele aus $\overline{\omega^-}$ bewirkt Inkonsistenz.
  Erweiterung von ω^- um Beispiele aus $\overline{\omega^+}$ bewirkt Inkonsistenz.
- *Nur die Erweiterung von (ω^+, ω^-) um ambige Beispiele $\boldsymbol{x} \in \omega^?$ ist zugleich **konsistent** und **produktiv**!*

Exploratives Lernen

Sukzessives Akquirieren produktiver neuer Beispiele, bis

1. der Versionenraum $\mathfrak{V}$ nur noch ein h enthält oder
2. der Versionenraum $\mathfrak{V}$ leergelaufen ist.

Lernen mit Orakel

(Algorithmus)

1. INITIALISIERUNG
   Setze $G \leftarrow \{\boldsymbol{\Omega}\}$ und $S \leftarrow \{\varnothing\}$ und $\omega^? \leftarrow \boldsymbol{\Omega}$.
2. EXPLORATIONSSCHRITT
   Solange $\omega^? \neq \varnothing$ gilt:
   a. Wähle ein Beispiel $\boldsymbol{x} \in \omega^?$ aus
   b. Befrage das Orakel nach $\boldsymbol{x} \in \mathcal{C}$
   c. Modifiziere den Versionenraum vermöge
      $$\mathfrak{V} \leftarrow \mathfrak{V}(\omega^+ \cup \{\boldsymbol{x}\}, \omega^-)$$
      im Fall einer positiven Antwort und vermöge
      $$\mathfrak{V} \leftarrow \mathfrak{V}(\omega^+, \omega^- \cup \{\boldsymbol{x}\})$$
      im Fall einer negativen Antwort des Orakels.
   d. Aktualisiere die Menge $\omega^?$ der ambigen Objekte
3. TERMINIERUNG
   Das Ergebnis ist h, falls $G = \{h\} = S$ gilt.

(Algorithmus)

Ambiguität und Rückweisung

Votierungstechniken für die Entscheidungsphase

Faules Lernen

Fallbasiertes Schließen

$$\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{\Omega}^+ & \boldsymbol{x} \in \omega^+ \\ \boldsymbol{\Omega}^- & \boldsymbol{x} \in \omega^- \\ \boldsymbol{\Omega}^? & \text{sonst} \end{cases}$$

(keine Verallgemeinerung)

Fleißiges Lernen

einer Hypothese $(\mathfrak{V} = \{h^*\})$

$$\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{\Omega}^+ & h^* \models \boldsymbol{x} \\ \boldsymbol{\Omega}^- & h^* \not\models \boldsymbol{x} \end{cases}$$

Orakel
Occam's Razor (MDL, BIC, AIC)
Wahrscheinlichkeiten ...

Einstimmigkeit

$$\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{\Omega}^+ & h \in \mathfrak{V}_S \Rightarrow h \models \boldsymbol{x} \\ \boldsymbol{\Omega}^- & h \in \mathfrak{V}_G \Rightarrow h \not\models \boldsymbol{x} \\ \boldsymbol{\Omega}^? & \text{sonst} \end{cases}$$

Generalkonsens

$$\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{\Omega}^+ & h \in \mathfrak{V}_G \Rightarrow h \models \boldsymbol{x} \\ \boldsymbol{\Omega}^- & \text{sonst} \end{cases}$$

Mehrheitsvotum

$$\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{\Omega}^+ & |\{h \in \mathfrak{V} \mid h \models \boldsymbol{x}\}| > |\mathfrak{V}|/2 \\ \boldsymbol{\Omega}^- & |\{h \in \mathfrak{V} \mid h \models \boldsymbol{x}\}| < |\mathfrak{V}|/2 \\ \boldsymbol{\Omega}^? & |\{h \in \mathfrak{V} \mid h \models \boldsymbol{x}\}| = |\mathfrak{V}|/2 \end{cases}$$

Gibbs-Sampling

Auswürfeln von $h^* \in \mathfrak{V}$ und

$$\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{\Omega}^+ & h^* \models \boldsymbol{x} \\ \boldsymbol{\Omega}^- & h^* \not\models \boldsymbol{x} \end{cases}$$

ILP — Induktive logische Programmierung

Hypothesenraumbias wird explizit durch eine logische Theorie $\mathcal{B}$ vorgegeben

Gegeben

Hypothesen $h \in \mathcal{H}$ sind **prädikatenlogische** Formeln
Objekte $\boldsymbol{x} \in \boldsymbol{\Omega}$ als Singleton-Hypothesen h_x
Positive und negative Lerndatensätze ω^+, ω^-
p.l. Formelmenge $\mathcal{B}$ als expliziter Bias („*Sachbereichstheorie*")

Gesucht

Eine Hypothese $h \in \mathcal{H}$ mit den Eigenschaften

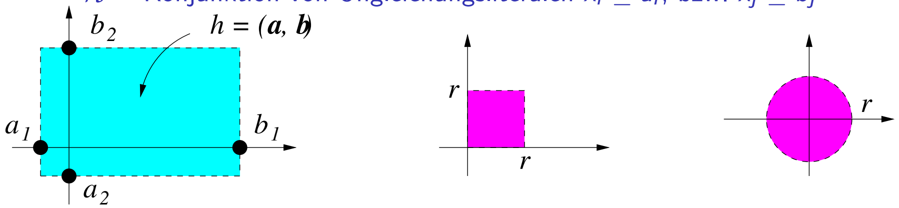
1. **Vollständigkeit** $\mathcal{B}, h \models \omega^+$
2. **Korrektheit** für alle $x \in \omega^-$ gilt $\mathcal{B}, h \not\models x$
3. **Konsistenz** $\mathcal{B}, h, \omega^+, \omega^- \not\models \texttt{false}$

die zudem ein Gütekriterium $\left\{ \begin{array}{c} \textit{speziell} \\ \textit{generell} \\ \textit{interessant} \\ \textit{kurz} \end{array} \right\}$ optimiert.

⇨ nicht entscheidbar in der Prädikatenlogik erster Stufe

Numerisches Beispiel $(\mathbf{\Omega} = \mathbb{R}^N)$

$\mathcal{H}$ = Konjunktion von Ungleichungsliteralen $x_i \geq a_i$, bzw. $x_j \leq b_j$



Hypothesenraum

$h \;=\; \{\mathbf{x} \mid a_i \leq x_i \leq b_i \text{ für alle } i\}$

Konzeptraum I

$C_r \;=\; \{\mathbf{x} \mid \max_i x_i \leq r\}$

Konzeptraum II

$C_r \;=\; \{\mathbf{x} \mid \|\mathbf{x}\| \leq r\}$

Computational Learning Theory

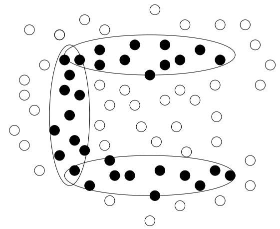
- Was wird (asymptotisch) gelernt? $\left\{\begin{array}{c}\text{korrektes Quadrat}\\ \text{u.U. nichts}\end{array}\right\}$
- Wie groß ist der erwartete Klassifikationsfehler? $\left\{\begin{array}{c}0\%\\ 22\%\end{array}\right\}$

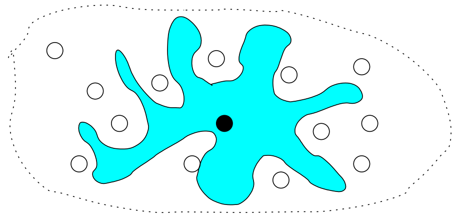
Michalskis Stern

Abgrenzung **eines** Positivbeispiels gegen **alle** Negativbeispiele

Problem

Die Menge ω^+ ist schwer gegen ω^- abgrenzbar.
ω^+ zerfällt jedoch in einfacher strukturierte Teilmengen.





Definition

Es seien $\omega^+ \subset \mathbf{\Omega}$, $\omega^- \subset \mathbf{\Omega}$ und $\mathbf{x} \in \omega^+$ ein Positivbeispiel. Die Hypothesenmenge

$$\mathcal{S}(\mathbf{x}|\omega^-) \stackrel{\text{def}}{=} \mathfrak{V}_G(\mathcal{H}, \{\mathbf{x}\}, \omega^-)$$

heißt **Stern** von $\mathbf{x}$ gegen ω^-.

Achtung!

Der Stern ist keine Hypothese, sondern ein Intervall.

Sterne und ihre Vereinigung

Lemma

Für jede Hypothese $h \in \mathcal{S}(\mathbf{x}|\omega^-)$ gilt:

1. *h wird von $\mathbf{x}$ erfüllt.* $\quad h \models \mathbf{x}$
2. *h wird von keinem $\mathbf{y} \in \omega^-$ erfüllt.* $\quad (\forall \mathbf{y} \in \omega^-)\; h \not\models \mathbf{y}$
3. *h ist maximal mit diesen Eigenschaften, d.h., es gilt:*

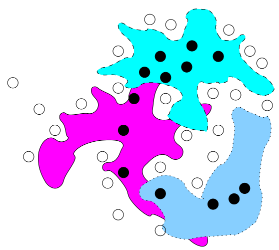
$$h' \supset h \;\Rightarrow\; h' \not\models \mathbf{x} \text{ oder ex. } \mathbf{y} \in \omega^- : \; h' \models \mathbf{y}$$

Lemma

*Aus **nichtleeren** Sternen lassen sich konsistente Disjunktionen konstruieren, d.h. die Vereinigungsmenge*

$$h^* \;=\; \bigcup_{x \in \omega^+} \mathcal{S}(x|\omega^-)$$

ist konsistent mit (ω^+, ω^-).



⊕ Disjunktion · ⊖ Hypothese

Sternerzeugungsalgorithmus

(Algorithmus)

1. Wähle zufällig ein $\mathbf{x}^* \in \omega^+$.
2. Erzeuge den Stern $\mathcal{S}(\mathbf{x}^*|\omega^-) = \mathfrak{V}_G(\mathcal{H}, \{\mathbf{x}^*\}, \omega^-)$.
3. Wähle eine Vorzugshypothese $h^* \in \mathcal{S}(\mathbf{x}^*|\omega^-)$ mit maximaler Präferenz $\gamma(h^*)$.
4. Wenn $h^* \models \mathbf{x}$ für alle $\mathbf{x} \in \omega^+$, so $\rightsquigarrow$ 6.
5. Tilge alle $\mathbf{x} \in \omega^+$ mit $h^* \models \mathbf{x}$ und $\rightsquigarrow$ 1.
6. Bilde die logische Disjunktion

$$h_{\text{dis}} \stackrel{\text{def}}{=} h_1 \vee \ldots \vee h_r$$

aller bislang erzeugten Hypothesen.

(Algorithmus)

Gegeben

$\mathcal{H}$, ω^+, ω^- und eine **Präferenzfunktion** $\gamma : \mathcal{H} \mapsto \mathbb{R}$

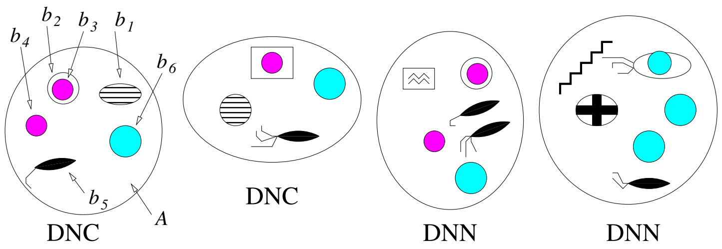
Gesucht

Eine disjunktive Beschreibung $h_1 \vee \ldots \vee h_r$, $h_i \in \mathcal{H}$, die konsistent mit (ω^+, ω^-) ist.

Beispiel — Konzeptualisierung von Krebszellen

Aufgabenstellung

Unterscheide Krebszellen (DNC) von gesunden Zellen (DNN) auf Grundlage numerischer, kategorialer und struktureller Zellmerkmale.



Mensch-Maschine-Mensch-Zyklus

(1) Definiere relevante Deskriptoren · etikettiere (ω^+, ω^-)
(2) Lerne induktiv passende Hypothesen für $\mathcal{C} = \mathcal{C}_{\mathrm{DNC}}$.
(3) Evaluiere, analysiere und modifiziere das Szenarium.

Beispiel — Konzeptualisierung von Krebszellen

Globale (zellbezogene) Merkmale

1. $circ \in \{1, 2, \ldots, 10\}$ (Anzahl der Zellsegmente)
2. $pplasm \in \{A, B, C, D\}$ (Protoplasmatyp der Zelle)

Lokale (segmentbezogene) Merkmale

- $shape(i) \in \{triangle, circle, ellipse, heptagon, square, boat, spring\}$
  (bzw. eine Baumstruktur dieser Formklassen)
- $texture(i) \in \{blank, shaded, black, grey, stripes, crossed, wavy\}$
- $weight(i) \in \{1, 2, 3, 4, 5\}$
- $orient(i) \in \{N, NE, E, SE, S, SW, W, NW\}$
- $contains(c, b_1, b_2, \ldots) \in \{T, F\}$
- $hastails(c, b_1, b_2, \ldots) \in \{T, F\}$

Beispiel — Konzeptualisierung von Krebszellen

Objektbeschreibung der ersten DNC-Zelle

$$
\begin{aligned}
& contains(c, b_1, \ldots, b_6) \wedge circ(c) = 8 \wedge pplasm(c) = A \\
& \wedge\ shape(b_1) = ellipse \wedge texture(b_1) = stripes \wedge weight(b_1) = 4 \\
& \wedge\ orient(b_1) = NW \wedge shape(b_2) = circle \wedge contains(b_2, b_3) \\
& \wedge\ texture(b_2) = blank \wedge weight(b_2) = 3 \wedge \ \ldots \\
& \wedge\ shape(b_6) = circle \wedge texture(b_6) = shaded \wedge weight(b_6) = 5
\end{aligned}
$$

DNC-Charakterisierung durch prädikatenlogische Formel

$$
\begin{aligned}
& \exists_1 b\ (weight(b) = 5) \\
& \exists_1 b\ (shape(b) = circle \wedge texture(b) = shaded \wedge weight(b) \geq 3) \\
& \exists b_1 \exists b_2\ (contains(b_1, b_2) \wedge shape(b_1) = circle \wedge shape(b_2) = circle) \\
& \ldots \wedge \ldots \wedge \ldots
\end{aligned}
$$

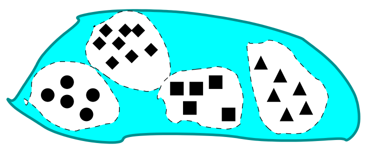
Konzeptuelle Klassifikation

Gegeben

Klassenspezifische Lernstichproben

$$\omega_\kappa \subseteq C_\kappa \subseteq \mathbf{\Omega}, \quad \kappa = 1, \ldots, K$$

für die Konzepte $C_1, \ldots, C_K \in \mathcal{C}$ mit $C_\kappa \cap C_\lambda = \varnothing$ für $\kappa \neq \lambda$.



Gesucht

Ein konsistentes System $h_1, \ldots, h_K \in \mathcal{H}$, d.h. für alle $1 \leq \kappa \leq K$ gilt:

$$\begin{cases} h_\kappa \models \boldsymbol{x} & \boldsymbol{x} \in \omega_\kappa \\ h_\kappa \not\models \boldsymbol{x} & \boldsymbol{x} \in \omega_\lambda,\ \lambda \neq \kappa \end{cases} \quad \text{und „}\textit{quodlibet}\text{“ sonst}$$

Versionenraum-Methode

Berechne für jede Objektklasse κ einen **diskriminativen VR**

$$\mathfrak{V}_\kappa = \mathfrak{V}(\mathcal{H}, \omega_\kappa, \bigcup_{\lambda \neq \kappa} \omega_\lambda)$$

Stern-Methode

Berechne für jedes κ eine **Sterndisjunktion**

$$h^\star_\kappa = \bigcup_{x \in \omega_\kappa} \mathcal{S}(x \mid \omega \setminus \omega_\kappa)$$

Votierung beim K-Klassen-Problem

$$\beta(\boldsymbol{x}) = (\beta_1, \ldots, \beta_K) \in \{1, 0, ?\}^K$$

Stern-Methode

Es gibt keine Fehlanzeigen.
Aber es gibt u.U. Konflikte.
Und es gibt u.U. Leerrunden.

Versionenraum-Methode

Es gibt Konflikte & Leerrunden.
Es gibt auch Fehlanzeigen:

$$\left\{ \begin{array}{c} \text{eine FA statt PRO} \\ \text{weniger als } K-1 \text{ CONs} \end{array} \right\}$$

Definition

Sei $\mathcal{A} \subseteq \boldsymbol{\Omega}$. Das Hypothesensystem $(h_1, \ldots, h_K)$ heißt **konzeptuelle Partition** von $\mathcal{A}$, wenn es für jedes $\boldsymbol{x} \in \mathcal{A}$ einen Klassenindex κ gibt mit

$$\forall \lambda = 1, \ldots, K : \qquad (h_\lambda \models \boldsymbol{x} \Leftrightarrow \lambda = \kappa)$$

Bemerkungen

1. Ein konsistentes System $(h_1, \ldots, h_K)$ ist konzeptuelle Partition seiner Lerndaten $\bigcup_\kappa \omega_\kappa$.
2. Läßt sich jedes konsistente System zu einer konzeptuellen Partition von $\boldsymbol{\Omega}$ erweitern?

Die Bayesregel

... ist der theoretisch optimale Klassifikator.

Satz

*Ist für den Objektraum $\mathcal{X}$ und das Klasseninventar $\mathcal{K} = \{1, \ldots, K\}$ die wahre Verbundverteilung $\mathrm{P}(\kappa, \boldsymbol{x})$ bekannt, so liefert die **Bayesentscheidungsregel** (MAP-Regel)*

$$\begin{aligned} \delta(\boldsymbol{x}) &= \underset{\kappa \in \mathcal{K}}{argmax}\, \mathrm{P}(\kappa|\boldsymbol{x}) \\ &= \underset{\kappa \in \mathcal{K}}{argmax}\, \frac{\mathrm{P}(\kappa) \cdot \mathrm{P}(\boldsymbol{x}|\kappa)}{\mathrm{P}(\boldsymbol{x})} \end{aligned}$$

die minimale erwartete Klassifikationsfehlerrate.

Marginal 1

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

Marginal 2

$$\mathrm{P}(\kappa) = \sum_{\boldsymbol{x} \in \mathcal{X}} \mathrm{P}(\kappa, \boldsymbol{x})$$

Bedingte Vtl.

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

Posterior Vtl.

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

Bemerkung

Die Aussage gilt natürlich nur, wenn das **korrekte** Wahrscheinlichkeitsmodell verwendet wird.

Die Bayesregel für diskrete Attribute

Kanonische multivariat-diskrete Verteilung („*Hypertabelle*“)

Lemma

*Der Objektraum $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_N$ enthalte ausschließlich **diskrete** Attribute mit Wertebereichen $\mathcal{X}_n$ der Größe $L_n = |\mathcal{X}_n|$.*

1. *Die gemeinsame Verteilung $\mathrm{P}(\kappa, \boldsymbol{x})$ ist durch die $K \cdot L_1 \cdot \ldots \cdot L_N$ Einträge*
$$p_{\kappa, x_1, \ldots, x_N} = \mathrm{P}(\kappa, \boldsymbol{x}), \qquad \kappa \in \mathcal{K},\ \boldsymbol{x} \in \mathcal{X}$$
eines $(1+N)$-dimensionalen Hyperwürfels $\boldsymbol{P} \in [0,1]^{K \times L_1 \times \ldots \times L_N}$ charakterisiert.
2. *Für einen etikettierten Lerndatensatz $\{(\kappa_t, \boldsymbol{x}_t) \mid t = 1..T\}$ mit den absoluten Häufigkeiten $T_{\kappa,\boldsymbol{x}}$, $(\kappa, \boldsymbol{x}) \in \mathcal{K} \times \mathcal{X}$, lauten die Maximum-Likelihood-Parameter*
$$\hat{b}_{\kappa, x_1, \ldots, x_N} = T_{\kappa, x_1, \ldots, x_N} \,/\, T \ .$$
3. *Die Bayesentscheidungsregel lautet $\delta(\boldsymbol{x}) = argmax_{\kappa \in \mathcal{K}}\, \hat{b}_{\kappa, x_1, \ldots, x_N}$.*

Beispiel — kanonische Bayesregel

... mit ML-geschätzten Verteilungsparametern

Lerndatensammlung „*Tennis*“

ω	*outlook*	*temp*	*humid*	*wind*	*Tennis?*
o_1	*sunny*	*hot*	*high*	*weak*	*no*
o_2	*sunny*	*hot*	*high*	*strong*	*no*
o_3	*overcast*	*hot*	*high*	*weak*	*yes*
o_4	*rain*	*mild*	*high*	*weak*	*yes*
o_5	*rain*	*cool*	*normal*	*weak*	*yes*
o_6	*rain*	*cool*	*normal*	*strong*	*no*
o_7	*overcast*	*cool*	*normal*	*strong*	*yes*
o_8	*sunny*	*mild*	*high*	*weak*	*no*
o_9	*sunny*	*cool*	*normal*	*weak*	*yes*
o_{10}	*rain*	*mild*	*normal*	*weak*	*yes*
o_{11}	*sunny*	*mild*	*normal*	*strong*	*yes*
o_{12}	*overcast*	*mild*	*high*	*strong*	*yes*
o_{13}	*overcast*	*hot*	*normal*	*weak*	*yes*
o_{14}	*rain*	*mild*	*high*	*strong*	*no*
o_{neu}	*sunny*	*cool*	*high*	*strong*	?

Parameter

$2 \cdot 3^2 \cdot 2^2 = 72$ Einträge

14 Einsen
58 Nullen

Neuzugang

Nulleintrag bei $(\textit{yes}, o_{\text{neu}})$ und $(\textit{no}, o_{\text{neu}})$.

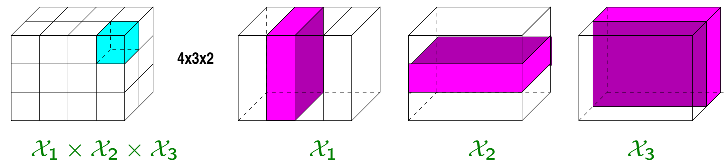
Nennerausdruck $\hat{\mathrm{P}}(o_{\text{neu}}) = 0$

Dann gilt für die a posteriori Wahrscheinlichkeit:

$$\mathrm{P}(\textit{no} \mid (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})^\top) \;=\; \text{undef.}$$

Die naive Bayesregel

Klassenbedingte statistische Unabhängigkeit zwischen allen Objektattributen



NBK-Entscheidungsregel

$$\delta(\boldsymbol{x}) \;=\; \operatorname*{argmax}_{\kappa \in \mathcal{K}} \mathrm{P}(\kappa, \boldsymbol{x}) \;=\; \operatorname*{argmax}_{\kappa \in \mathcal{K}} \left\{ \mathrm{P}(\kappa) \cdot \prod_{n=1}^{N} \mathrm{P}(x_n | \kappa) \right\}$$

(maximale faktorisierte Verbundwahrscheinlichkeit)

Modellparameter und ihre ML-Schätzwerte

$$\hat{a}_\kappa \;=\; \frac{T_\kappa}{T}\,, \qquad \hat{b}_{\xi|\kappa,n} \;=\; \frac{T_{\kappa,n,\xi}}{T_\kappa}\,, \qquad T_\kappa \;=\; \sum_{\xi \in \mathcal{X}_1} T_{\kappa,1,\xi}\,, \qquad \begin{cases} \kappa = 1..K \\ n = 1..N \\ \xi = 1..L_n \end{cases}$$

Das sind $K \cdot \sum_n L_n$ Parameter statt $K \cdot \prod_n L_n$ Parameter!

Beispiel — naive Bayesregel

... mit ML-geschätzten Verteilungsparametern

Attribut „*outlook*“

	sunny	*over*	*rain*	Σ
yes	2	4	3	9
no	3	0	2	5
Σ	5	4	5	14

Attribut „*humidity*“

	high	*normal*	Σ
yes	3	6	9
no	4	1	5
Σ	7	7	14

Attribut „*temp*“

	hot	*mild*	*cool*	Σ
yes	2	4	3	9
no	2	2	1	5
Σ	4	6	4	14

Attribut „*wind*“

	weak	*strong*	Σ
yes	6	3	9
no	2	3	5
Σ	8	6	14

Parametertabelle und Neuklassifikation (6 + 6 + 4 + 4 = 20 Einträge)

$$\mathrm{P}(\textit{no}, \textit{sunny}, \textit{cool}, \textit{high}, \textit{strong}) \;=\; \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \;=\; \frac{180}{8750} \;=\; 0.02057$$

$$\mathrm{P}(\textit{yes}, \textit{sunny}, \textit{cool}, \textit{high}, \textit{strong}) \;=\; \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \;=\; \frac{486}{91854} \;=\; 0.005291$$

$$\mathrm{P}(\textit{no} \mid (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})^\top) \;=\; \frac{0.02057}{0.02057 + 0.005291} \;=\; 0.7954$$

NTF — Nichtnegative Tensorfaktorisierung

Mischung naiver Verbundverteilungen von $N \in \{2, 3\}$ nominalen Attributen

Matrix $(\boldsymbol{\Omega}_1 \times \boldsymbol{\Omega}_2)$

Verteilungsparameter

$$\mathrm{P}(i, j) \;=:\; x_{ij}$$

Naive Faktorisierung

$$\mathrm{P}(i, j) \;=\; p_1(i) \cdot p_2(j)$$

Mischungsmodell

$$\mathrm{P}(i, j) \;=\; \sum_{m=1}^{M} \underbrace{\pi_m \cdot p_1^{(m)}(i)}_{v_{im}} \cdot \underbrace{p_2^{(m)}(j)}_{a_{jm}}$$

Reduktion

$L_1 \cdot L_2$ ⇨ $M \cdot (1 + L_1 + L_2)$

Würfel $(\boldsymbol{\Omega}_1 \times \boldsymbol{\Omega}_2 \times \boldsymbol{\Omega}_3)$

Verteilungsparameter

$$\mathrm{P}(i, j, k) \;=:\; x_{ijk}$$

Naive Faktorisierung

$$\mathrm{P}(i, j, k) \;=\; p_1(i) \cdot p_2(j) \cdot p_3(k)$$

Mischungsmodell

$$\mathrm{P}(i, j, k) \;=\; \sum_{m=1}^{M} \pi_m \cdot p_1^{(m)}(i) \cdot p_2^{(m)}(j) \cdot p_3^{(m)}(k)$$

Reduktion

$L_1 \cdot L_2 \cdot L_3$ ⇨ $M \cdot (1 + L_1 + L_2 + L_3)$

NTF — Nichtnegative Tensorfaktorisierung

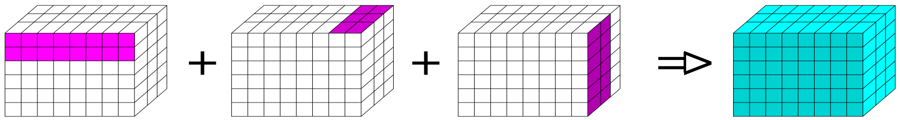
Mischung naiver Verbundverteilungen nominaler Attribute

Wahrscheinlichkeitshyperwürfel ($\mathbf{\Omega}_1 \times \mathbf{\Omega}_2 \times \ldots \times \mathbf{\Omega}_N$)

Naive Mischung

$$\mathrm{P}(x_1, \ldots, x_N) \;=\; \sum_{m=1}^{M} \pi_m \cdot \prod_{n=1}^{N} p_n^{(m)}(x_n)$$

Parameter lernen nach EM-Prinzip *(expectation-maximization)*



Reduktion

$L_1 \cdot \ldots \cdot L_N$ ⇨ $M \cdot (1 + L_1 + \ldots + L_N)$

$\prod \rightsquigarrow M \cdot \sum$

EM-Algorithmus für das NTF-Modell

$\boxed{1}$ Initialisierung · $\boxed{2}$ E-Schritt · $\boxed{3}$ M-Schritt · $\boxed{4}$ Abbruch

A posteriori Wahrscheinlichkeiten der Komponentenauswahl

Für jedes Lerndatensatzobjekt $\mathbf{x}_1, \ldots, \mathbf{x}_T$ berechne

$$\gamma_t(m) \;\stackrel{\text{def}}{=}\; \mathrm{P}(\mathbb{M} = m \mid \mathbf{x}_t, \boldsymbol{\theta}^{\text{alt}}) \;=\; \pi_m^{\text{alt}} \cdot \prod_{n=1}^{N} \theta^{\text{alt}}_{m,n,x_{tn}} \,/\, \mathrm{P}^{\text{alt}}(\mathbf{x}_t)$$

Neuschätzung durch a posteriori Erwartungswerte

$$\hat{\pi}_m \;=\; \sum_{t=1}^{T} \gamma_t(m) \,/\, T \quad \text{und} \quad \hat{\theta}_{m,n,\xi} \;=\; \sum_{t=1}^{T} \gamma_t(m) \cdot \mathbf{I}_{x_{tn}=\xi} \Big/ \sum_{t=1}^{T} \gamma_t(m)$$

Startparameter

zufällig · wiederholt · lokale Optima

Rechenaufwand

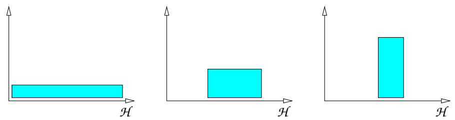
$\mathrm{O}(I_{\max} \cdot T \cdot M \cdot (N + \sum_n L_n))$

Parametrische (modellgetriebene) Bayesregel

Optimale Entscheidungsregel unter der Annahme $C^* \in \mathcal{H} \neq \mathcal{C}$

Problem

Die Auswahl und exklusive Nutzung der bestpassenden Hypothese aus $\mathcal{H}$ ist willkürlich und wegen $|\omega| < \infty$ suboptimal.



$\mathrm{P}(h)$ $\quad$ $\mathrm{P}(h|\mathbf{x}_1)$ $\quad$ $\mathrm{P}(h \mid \mathbf{x}_1, \mathbf{x}_2)$

Bayestheorie

1. Hypothesen sind nicht nur wahr oder falsch
2. Hypothesen treffen „weiche“ Entscheidungen
3. Jedes Lernbeispiel erhöht/vermindert inkrementell die Hypothesenwahrscheinlichkeiten
4. Vorwissen läßt sich mit den Lerndaten verzahnen
5. Mathematisch abgesicherter Votierungsmechanismus

Hypothesen- und Objektwahrscheinlichkeiten

Lemma (A priori und a posteriori Hypothesenwahrsch'keit)

*Für alle Hypothesen $h \in \mathcal{H}$ und gegebene Daten $\omega \subset \mathbf{\Omega}$ gilt die folgende Darstellung der **a posteriori** Hypothesenwahrscheinlichkeit:*

$$\mathrm{P}(h|\omega) \;=\; \frac{\mathrm{P}(\omega|h) \cdot \mathrm{P}(h)}{\mathrm{P}(\omega)}$$

„Deduktion“ ⇧ $\mathrm{P}(h)$ · *„Abduktion“* ⇧ $\mathrm{P}(\omega|h)$ · *„explaining-away“* ⇩ $\mathrm{P}(\omega)$

Definition

Für ein K-Klassenproblem über $\mathbf{\Omega}$ mit Hypothesenraum $\mathcal{H}$ und den Lerndaten ω heißt

$$\delta_{\text{Bayes}}(\mathbf{x}) \;=\; \operatorname*{argmax}_{\kappa} \sum_{h \in \mathcal{H}} \mathrm{P}(\mathbf{x} \in C_\kappa \mid h) \cdot \mathrm{P}(h|\omega)$$

die **Bayes-Entscheidungsregel** für das Objekt $\mathbf{x}$.

Bemerkung

Diese Entscheidungsregel realisiert die **minimale asymptotische Fehlerrate**.

Hypothesenauswahltechniken

MDL-Prinzip

Minimum Description Length (Rissanen'87)

$$
\begin{aligned}
h_{\mathrm{MAP}} &= \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{P}(\omega|h) \cdot \mathrm{P}(h) \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} \{-\log_2 \mathrm{P}(\omega|h) - \log_2 \mathrm{P}(h)\}
\end{aligned}
$$

(Algorithmus)

1. Wähle eine Codierung $\mathfrak{C}_1$ für die Hypothesen
2. Wähle eine bedingte Codierung $\mathfrak{C}_2$ für die Klassen
3. Berechne die „*kürzeste Erklärung*"

$$h_{\mathrm{MDL}} \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{h \in \mathcal{H}} \{\ell_{\mathfrak{C}_1}(h) + \ell_{\mathfrak{C}_2}(\omega|h)\}$$

der Lerndaten.

ML-Schätzer

$h_{\mathrm{ML}} = \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{P}(\omega|h)$

MAP-Schätzer

$h_{\mathrm{MAP}} = \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{P}(h|\omega)$

Gibbs-Sampler

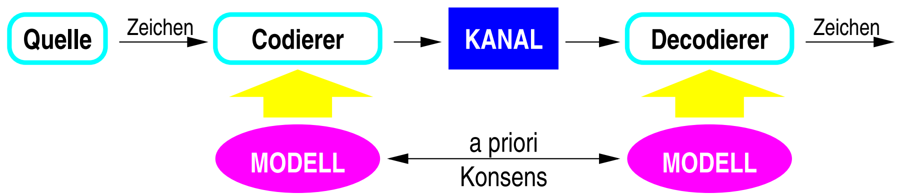
$h_{\mathrm{GS}} \sim \mathrm{P}(h|\omega)$

Versionenraum

$h_{\mathrm{VR}} \in \mathfrak{V}(\mathcal{H}, \omega)$

Noiseless Source Coding Theorem

Quellenentropie $\hat{=}$ minimale Bitanzahl nach optimaler Datenkompression



Satz (Shannon 1949)

Ein Zufallsprozeß erzeuge Zeichenfolgen über dem Alphabet $\{s_1, \ldots, s_L\}$ mit den Wahrscheinlichkeiten $q_1, \ldots, q_L$.

1. *Die **optimale** Codierung dieser Quelle verwendet für jedes Zeichen s_l ein Codewort der Länge $-\log_2 q_l$ Bit.*
2. *Ihre mittlere Codewortlänge beträgt $\mathcal{H}(q_1, \ldots, q_L)$ Bit.*

Bayesregel für gemischte Attributskalen

$$\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{\boldsymbol{\Omega}' = \mathbb{R}^{N'}} \times \underbrace{\mathcal{X}_1 \times \ldots \times \mathcal{X}_{N''}}_{\boldsymbol{\Omega}''}$$

Normalzerlegung

$$\mathrm{P}(\boldsymbol{z}) = \mathrm{P}(\boldsymbol{y}) \cdot \mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \mathrm{P}(y_1, \ldots, y_{N''}) \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{S}_{\boldsymbol{y}})$$

- $L^{\times} = \prod_{n=1}^{N''} L_n$ kanonische W'keitsparameter
- $N' + \binom{N'}{2}$ Dichteparameter je NV-Dichte
- ➪ insgesamt $\mathrm{O}(K \cdot L^{\times} \cdot N'^2)$ Parameter
- Unabhängigkeitsannahme für $\boldsymbol{\Omega}''$ bringt wenig Vorteile.
- Unabhängigkeit in $\boldsymbol{\Omega}'$ reduziert auf $\mathrm{O}(K \cdot L^{\times} \cdot N')$ Parameter.

Diskriminative Klassifikatoren

$\kappa(\boldsymbol{x}) = \operatorname{argmax}_\lambda h_\lambda(\boldsymbol{x})$

Definition

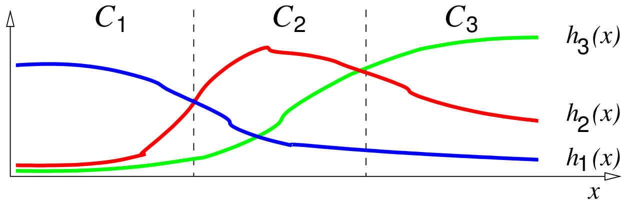
Es sei $C_1, \ldots, C_K$ ein K-Klassen-Problem über $\boldsymbol{\Omega} = \mathbb{R}^N$. Die Elemente von $\boldsymbol{h} = (h_1, \ldots, h_K)^\top$ mit

$$h_\kappa : \mathbb{R}^N \to \mathbb{R} \;, \quad \kappa = 1, \ldots, K$$

heißen **Trennfunktionen** der Klassen $\kappa = 1, \ldots, K$.
Die Abbildungen $\boldsymbol{d} = (d_1, \ldots, d_K)^\top$ mit

$$d_\kappa : \boldsymbol{x} \mapsto \begin{cases} 1 & \boldsymbol{x} \in C_\kappa \\ 0 & \boldsymbol{x} \notin C_\kappa \end{cases}$$

heißen **ideale Trennfunktionen** des Problems.



Quadratmittelklassifikator

Willkürlicher Zielausdruck — willkürliches Straffunktional

Definition

Es sei $C_1, \ldots, C_K$ ein K-Klassen-Problem über $\boldsymbol{\Omega} = \mathbb{R}^N$ und $\mathcal{H}$ eine Menge von Trennfunktionen. Die Trennfunktion $\boldsymbol{h} \in \mathcal{H}$ mit minimalem erwarteten quadratischen Fehler

$$\varepsilon(\boldsymbol{h}) \overset{\text{def}}{=} \mathcal{E}[\|\boldsymbol{h}(\mathbb{X}) - \boldsymbol{d}(\mathbb{X})\|^2]$$

heißt **Quadratmitteldiskriminante**, der zugehörige Klassifikator heißt **Quadratmittelklassifikator**.

Sind ferner die Lerndaten $\omega_1, \ldots, \omega_K$ gegeben, so heißt der Klassifikator mit minimalem Fehler

$$\varepsilon(\boldsymbol{h}, \{\omega_\kappa\}) \overset{\text{def}}{=} \sum_{\kappa=1}^{K} \sum_{\boldsymbol{x} \in \omega_\kappa} \left\| \boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{e}^{(\kappa)} \right\|^2$$

empirischer QMK. Dabei bezeichne $\boldsymbol{e}^{(\kappa)}$ den κ-ten Einheitsvektor.

Lernen als skalare Regressionsaufgabe

Zerlegung in Zweiklassenprobleme

Für jedes κ ergibt sich das QM-Approximationsproblem

$$h_\kappa \quad \approx \quad d_\kappa : \boldsymbol{x} \mapsto \begin{cases} 1 & \boldsymbol{x} \in \omega^+, \ \omega^+ = \omega_\kappa \\ 0 & \boldsymbol{x} \in \omega^-, \ \omega^- = \bigcup_{\lambda \neq \kappa} \omega_\lambda \end{cases}$$

Skalares Regressionsproblem

Für die Daten $\left\{(\boldsymbol{x}_t, y_t) \in \mathbb{R}^N \times \mathbb{R} \mid t = 1, \ldots, T\right\}$ finde die Regressionsfunktion $h \in \mathcal{H}$ mit minimalem Fehler

$$\varepsilon(h) \overset{\text{def}}{=} \sum_{t=1}^{T} (y_t - h(\boldsymbol{x}_t))^2$$

$$\left\{ \begin{matrix} \text{quadratisches Fehlermaß} \\ \{0,1\}\text{-Zielgröße} \end{matrix} \right\} \Rightarrow \left\{ \begin{matrix} \text{entkoppelte Diskriminantfunktionen} \\ \text{kein Problem mit } K > 2 \end{matrix} \right\}$$

Multivariate lineare Regression

Linearer Ansatz

$$h(\boldsymbol{x}) \;=\; \sum_n a_n \cdot x_n \;=\; \boldsymbol{a}^\top \boldsymbol{x} \;, \qquad \boldsymbol{x}, \boldsymbol{a} \in \mathbb{R}^N$$

Affiner Ansatz

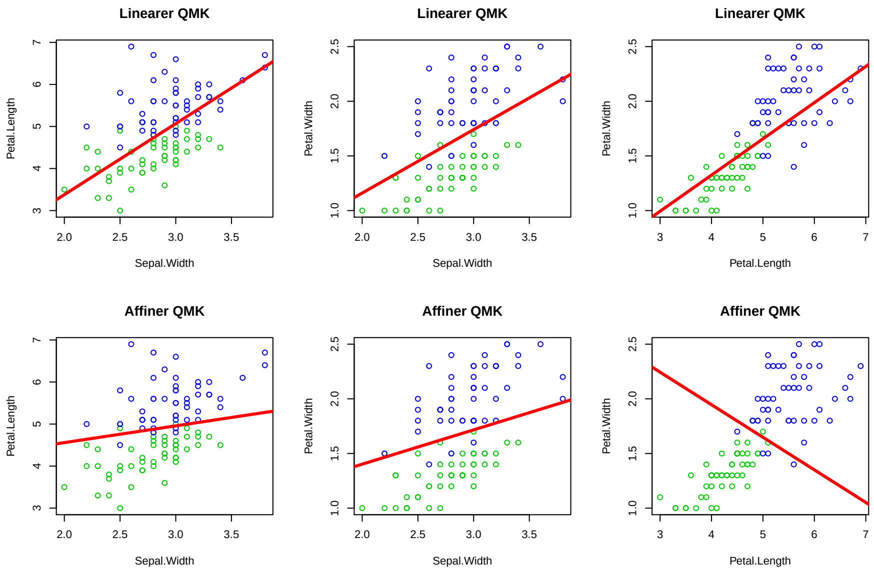
$$h(\boldsymbol{x}) \;=\; a_0 + \sum_n a_n \cdot x_n \;=\; \boldsymbol{a}^\top \boldsymbol{x}' \;, \qquad \begin{cases} \boldsymbol{a} \in \mathbb{R}^{N+1} \\ \boldsymbol{x}' \overset{\text{def}}{=} (1, \boldsymbol{x}^\top)^\top \end{cases}$$

Was heißt hier eigentlich „Regressionsproblem“ ?

- Datenmodell $\mathbb{Y} = h(\mathbb{X}_1, \ldots, \mathbb{X}_N) + \mathbb{E}$ mit Störterm $\mathbb{E} \sim \mathcal{N}(0, \sigma^2)$
- Datenprobe $\{\boldsymbol{x}_t, y_t\}_1^T$ bzw. gemeinsame Datenverteilung $f_{\mathbb{X},\mathbb{Y}}(\cdot,\cdot)$
- ⇨ Posterior-Erwartungswerte $\hat{h}(\boldsymbol{x}) = \mathcal{E}_{\mathbb{Y}|\boldsymbol{x}}[\mathbb{Y}]$, also $\hat{h}(\boldsymbol{x}) = \int f(y|\boldsymbol{x}) \cdot y \, dy$

Linearer vs. affiner Quadratmittelklassifikator

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



Linearer versus affiner Ansatz

Lineare Funktionen allein beschreiben wegen $h(\mathbf{0}) = 0$ ausschließlich Hyperebenen, die durch den Koordinatenursprung verlaufen und sind als Regressionsmodell unzureichend. Affine Funktionen verfügen zusätzlich über den y-Schnittpunkt a_0 (*intercept*); affine Regression kann aber leicht auf lineare Regression zurückgeführt werden. Wir verwenden wieder die Notation $\boldsymbol{X}$, $\boldsymbol{y}$ für Datenmatrix und Zielwertevektor und betrachten den Abweichungsvektor $\boldsymbol{y} - a_0\mathbf{1} - \boldsymbol{X}\boldsymbol{a}$ des affinen Modells sowie den resultierenden quadratischen Fehler:

$$\begin{aligned}
\varepsilon(a_0, \boldsymbol{a}) &= \|\boldsymbol{y} - a_0\mathbf{1} - \boldsymbol{X}\boldsymbol{a}\|^2 \\
&= (\boldsymbol{y} - a_0\mathbf{1} - \boldsymbol{X}\boldsymbol{a})^\top \cdot (\boldsymbol{y} - a_0\mathbf{1} - \boldsymbol{X}\boldsymbol{a}) \\
&= (\boldsymbol{y}^\top - a_0\mathbf{1}^\top - \boldsymbol{a}^\top\boldsymbol{X}^\top) \cdot (\boldsymbol{y} - a_0\mathbf{1} - \boldsymbol{X}\boldsymbol{a}) \\
&= \underbrace{\|\boldsymbol{y}\|^2}_{c} + \underbrace{a_0^2 T - 2a_0 T\mu_y}_{\varepsilon(a_0)} + \underbrace{\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} - 2\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{y}}_{\varepsilon(\boldsymbol{a})} + \underbrace{2a_0 T\boldsymbol{\mu}_x^\top\boldsymbol{a}}_{0}
\end{aligned}$$

Wenn wir o.B.d.A. mittelwertfreie Vektordaten annehmen ($\boldsymbol{\mu}_x = \mathbf{0}$), so verschwindet der Kopplungsterm und wir dürfen a_0 und $\boldsymbol{a}$ separat optimieren. Für a_0 ergibt sich nach Nullsetzen der Ableitung

$$\partial\varepsilon(a_0) \,/\, \partial a_0 = 2Ta_0 - 2T\mu_y$$

der Minimalwert $a_0 = \mu_y$. Der Fehler $\varepsilon(\boldsymbol{a})$ und die Konstante $\|\boldsymbol{y}\|^2$ ergeben zusammen den Minimierungsausdruck der linearen Regressionsaufgabe ...

Multivariate lineare Regression

Lösen des Systems der Gaußschen Normalengleichungen

Satz

Es seien die Regressionsdaten $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^N \times \mathbb{R}$, $t = 1, \ldots, T$ in der Matrixnotation

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^\top , \quad \boldsymbol{y} = (y_1, \ldots, y_T)^\top$$

dargestellt, und es sei $h : \boldsymbol{x} \mapsto \boldsymbol{a}^\top\boldsymbol{x}$ linear. Dann lautet der quadratische Regressionsfehler

$$\varepsilon(h) = \varepsilon(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2$$

und wird durch jede Lösung $\boldsymbol{a}$ der ***Gaußschen Normalengleichungen***

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = \boldsymbol{X}^\top\boldsymbol{y}$$

minimiert.

Beweis.

Zur Lösung des Quadratmittelproblems setzen wir die partiellen Ableitungen der Koeffizienten $a_1, \ldots, a_N$ gleich Null — wir verwenden die Gradientenvektorschreibweise:

$$\begin{aligned}
\nabla_{\boldsymbol{a}}\varepsilon(\boldsymbol{a}) &= \nabla_{\boldsymbol{a}} \left\{ \|\boldsymbol{y}\|^2 + \boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{a} \right\} \\
&= \mathbf{0} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} - 2\boldsymbol{X}^\top\boldsymbol{y} \\
&= 2 \cdot \left(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} - \boldsymbol{X}^\top\boldsymbol{y} \right) \\
&\overset{!}{=} \mathbf{0}
\end{aligned}$$

□

Bemerkung

Das LGS $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = \boldsymbol{X}^\top\boldsymbol{y}$ heißt System der *Gaußschen Normalengleichungen*. Wir schreiben auch kürzer $\boldsymbol{R}\boldsymbol{a} = \boldsymbol{m}$; dabei ist $\boldsymbol{R}$ wieder die unzentrierte, unnormierte Kovarianzmatrix der Vektordaten.

Ausgleichsrechnung und Lineare Gleichungssysteme

Was Sie schon immer über lineare Algebra wissen wollten, aber nie zu fragen wagten

$$\boldsymbol{X} \cdot \boldsymbol{a} \overset{!}{=} \boldsymbol{y}$$ mit dem Fehlervektor $\boldsymbol{e} := \boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}$

LGS eindeutig

Matrix $\boldsymbol{X}$ ist quadratisch und vollrangig.

$f : \boldsymbol{z} \mapsto \boldsymbol{X}\boldsymbol{z}$ bijektiv

$\boldsymbol{a} = \boldsymbol{X}^{-1}\boldsymbol{y}$ ist die eindeutige Lösung mit Fehler $\boldsymbol{e} = \boldsymbol{0}$.

… überbestimmt

Matrix $\boldsymbol{X}$ hat den vollen Spaltenrang.

$f : \boldsymbol{z} \mapsto \boldsymbol{X}\boldsymbol{z}$ ist injektiv
$\boldsymbol{X}^\top\boldsymbol{X}$ ist regulär

$\boldsymbol{a} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1} \cdot \boldsymbol{X}^\top\boldsymbol{y}$ ist eine Lösung mit minimalem Gesamtfehler $\|\boldsymbol{e}\|$.

… unterbestimmt

Matrix $\boldsymbol{X}$ hat den vollen Zeilenrang.

$f : \boldsymbol{z} \mapsto \boldsymbol{X}\boldsymbol{z}$ ist surjektiv
$\boldsymbol{X}\boldsymbol{X}^\top$ ist regulär

$\boldsymbol{a} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1} \cdot \boldsymbol{y}$ ist eine Lösung mit Fehler $\boldsymbol{e} = \boldsymbol{0}$ und minimaler Länge $\|\boldsymbol{a}\|$.

Bemerkung
Das Gaußsche Normalengleichungssystem ist entweder eindeutig lösbar oder besitzt unendlich viele Lösungen.

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Minimalnormlösung des GNG-Systems

Lemma

Es sei $\boldsymbol{R} \cdot \boldsymbol{a} = \boldsymbol{m}$ bzw. $\boldsymbol{X}^\top\boldsymbol{X} \cdot \boldsymbol{a} = \boldsymbol{X}^\top\boldsymbol{y}$ das GNG-System einer linearen Regressionsaufgabe.

1. *Die Matrix $\boldsymbol{R}$ ist symmetrisch und positiv-semidefinit.*
2. *Das Gleichungssystem hat stets mindestens eine Lösung.*
3. *Ist $\boldsymbol{R}$ invertierbar, so existiert eine eindeutige Lösung:*

$$\boldsymbol{a}^* = \boldsymbol{R}^{-1} \cdot \boldsymbol{m}$$

4. *Ist $\boldsymbol{X}^+$ die Pseudoinverse der Datenmatrix, so löst*

$$\boldsymbol{a}^+ = \boldsymbol{X}^+ \cdot \boldsymbol{y}$$

das Gleichungssystem und besitzt unter allen Lösungen die minimale Norm $\|\boldsymbol{a}^+\|$.

Die Berechnung der Minimalnormlösung ist **nicht praktikabel**!

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

System der Gaußschen Normalengleichungen

Linearer Quadratmittelklassifikator $(K = 2)$

$$\begin{aligned}
\boldsymbol{R} \cdot \boldsymbol{a} &= \boldsymbol{m} \\
\boldsymbol{R} &= {}^1\!/_T \cdot \boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
\boldsymbol{m} &= {}^1\!/_T \cdot \boldsymbol{X}^\top\boldsymbol{y} = {}^1\!/_T \cdot \sum_{\omega^+} \boldsymbol{x}_t = p^+ \cdot \boldsymbol{\mu}^+
\end{aligned}$$

$p^+ = {}^{T^+}\!/_T$, $\boldsymbol{\mu}^+$ Positivstatistiken; $\boldsymbol{R}$ Momentenmatrix der Gesamtprobe.

Linearer Quadratmittelklassifikator $(K > 2)$

$$\begin{aligned}
\boldsymbol{R} \cdot \boldsymbol{a}_1 &= \boldsymbol{m}_1 \\
\vdots &= \vdots \qquad , \qquad \boldsymbol{m}_\kappa = \frac{1}{T} \cdot \sum_{\omega_\kappa} \boldsymbol{x}_t = p_\kappa \cdot \boldsymbol{\mu}_\kappa \\
\boldsymbol{R} \cdot \boldsymbol{a}_K &= \boldsymbol{m}_K
\end{aligned}$$

Kompaktschreibweise: $\boldsymbol{R} \cdot \boldsymbol{A} = \boldsymbol{M}$ mit $\boldsymbol{M} = (p_1\boldsymbol{\mu}_1, \ldots, p_K\boldsymbol{\mu}_K)$.

Beweis.

Für eine beliebige Rechteckmatrix mit der SV-Zerlegung $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top$ heißt die Matrix

$$\boldsymbol{X}^+ = \boldsymbol{U}\boldsymbol{D}^+\boldsymbol{V}^\top$$

die **Moore-Penrose-Inverse** oder **Pseudoinverse**. Die Pseudoinverse $\boldsymbol{D}^+$ einer Diagonalmatrix $\boldsymbol{D}$ wiederum enthält auf ihrer Diagonalen die Pseudo-Reziproken:

$$d_n^+ = \begin{cases} 1/d_n & d_n \neq 0 \\ 0 & d_n = 0 \end{cases} , \qquad n = 1, \ldots, N$$

Diese Pseudoinverse gehorcht der **Moore-Penrose-Gleichung**, denn es gilt:

$$\begin{aligned}
\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^+ &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \cdot \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \cdot \boldsymbol{U}\boldsymbol{D}^+\boldsymbol{V}^\top \\
&= \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{D}^+\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{X}^\top
\end{aligned}$$

Folglich löst $\boldsymbol{a}^+ = \boldsymbol{X}^+\boldsymbol{y}$ auch die Gaußschen Normalengleichungen:

$$T \cdot \boldsymbol{R}\boldsymbol{a}^+ = \boldsymbol{X}^\top\boldsymbol{X} \cdot \boldsymbol{X}^+\boldsymbol{y} = \boldsymbol{X}^\top \cdot \boldsymbol{y} = T \cdot \boldsymbol{m}$$

Der Beweis der Minimaleigenschaft erfordert einen Lagrange-Ansatz:

$$\frac{1}{2} \cdot \|\boldsymbol{a}\|^2 + \lambda \cdot \left\|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} - \boldsymbol{X}^\top\boldsymbol{y}\right\|^2 \overset{!}{\to} \text{MIN}$$

□

Gratregularisierung

Lemma

Der regularisierte quadratische Regressionsfehler

$$\varepsilon_\lambda(\boldsymbol{a}) \;=\; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2 + \lambda \cdot \|\boldsymbol{a}\|^2$$

$(\lambda > 0)$ *wird durch die (eindeutige) Lösung*

$$\boldsymbol{a}_\lambda^* \;=\; (\boldsymbol{R}_\lambda)^{-1} \cdot \boldsymbol{m}\,, \qquad \boldsymbol{R}_\lambda \stackrel{def}{=} \boldsymbol{R} + \lambda \boldsymbol{E}$$

minimiert.

Beweis.

Der regularisierte Quadratmittelfehler besitzt den Gradientenvektor

$$\nabla_{\boldsymbol{a}}\varepsilon_\lambda(\boldsymbol{a}) \;=\; \nabla_{\boldsymbol{a}}\varepsilon(\boldsymbol{a}) + \lambda \cdot \nabla_{\boldsymbol{a}}\|\boldsymbol{a}\|^2 \;=\; 2 \cdot (\boldsymbol{R}\boldsymbol{a} + \lambda\boldsymbol{a} - \boldsymbol{m}) \;=\; 2 \cdot (\boldsymbol{R}_\lambda \boldsymbol{a} - \boldsymbol{m})$$

Die Gratregularisierungsmatrix ist für $\lambda \neq 0$ stets invertierbar, denn wegen

$$\boldsymbol{R}_\lambda \;=\; \boldsymbol{R} + \lambda\boldsymbol{E} \;=\; \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top + \lambda\boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^\top \;=\; \boldsymbol{U}\cdot(\boldsymbol{D}^2 + \lambda\boldsymbol{E})\cdot\boldsymbol{U}^\top \;=\; \boldsymbol{U}\cdot(\boldsymbol{D}^2)_\lambda \cdot \boldsymbol{U}^\top$$

besitzen alle Eigenwerte von $\boldsymbol{R}_\lambda$ die Form $d_n^2 + \lambda > 0$. □

Gewichtete & nichtquadratische Regression

Historische Wurzeln des IRLS: *„Iteratively Reweighted Least Squares"*

Quadratischer Fehler	$\sum_t (\boldsymbol{x}_t^\top \boldsymbol{a} - y_t)^2$	$=$	$\|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}\|_2^2$
Allgemeiner L_p-Fehler	$\sum_t \lvert\boldsymbol{x}_t^\top \boldsymbol{a} - y_t\rvert^p$	$=$	$\|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}\|_p^p$
Gewichteter Fehler	$\sum_t w_t^2 \cdot (\boldsymbol{x}_t^\top \boldsymbol{a} - y_t)^2$	$=$	$\|\boldsymbol{W}\cdot(\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y})\|_2^2$

$\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_T)$

Gewichtete Ausgleichsrechnung

Wegen $\|\boldsymbol{W}\cdot(\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y})\|_2^2 \;=\; \|\boldsymbol{W}\boldsymbol{X}\boldsymbol{a} - \boldsymbol{W}\boldsymbol{y}\|_2^2 \;=\; \left\|\tilde{\boldsymbol{X}}\boldsymbol{a} - \tilde{\boldsymbol{y}}\right\|_2^2$ lautet der Lösungskoeffizientenvektor $\boldsymbol{a} \;=\; (\boldsymbol{X}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\cdot\boldsymbol{X}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{y}$

Ausgleichsrechnung in der L_p-Fehlernorm (Betrag/Minimum)

Die Fehlerminimierung kann wegen

$$\|\boldsymbol{e}\|_p^p \;=\; \sum_t |e_t|^p \;=\; \sum_t |e_t|^{p-2}|e_t|^2 \;=\; \sum_t w_t^2 e_t^2$$

auf IRLS mit Gewichten $w_t = |e_t|^{(p-2)/2}$ zurückgeführt werden.

Lineare *versus* nichtlineare QMK

Angriffspunkt: 1.Quellvariable 2.Berechnungsweg 3.Zielvariable

Termexpansion

GNGS für alle Koeffizienten

- Linear & affin $\mathrm{O}(N^1)$
- Quadratisch $\mathrm{O}(N^2)$
- Kubisch $\mathrm{O}(N^3)$
- Polynomansatz $\mathrm{O}(\binom{N+p}{p})$

Nominale Attribute

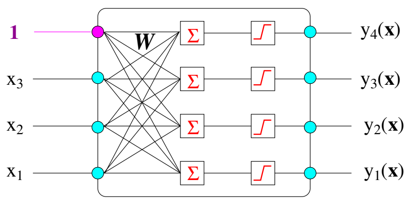
Kontrastmatrizen $(L_n - 1)$

- ohne Interaktionsterme $\mathrm{O}(\ell),\ \ell = \sum_n L_n$
- einfache Interaktionsterme $\mathrm{O}(\ell^2)$

Neuronale Berechnungsmodelle

Error Backpropagation

- Mehrschichtenperzeptron
- Radiale Basisfunktionen
- Time-Delay Neural Network

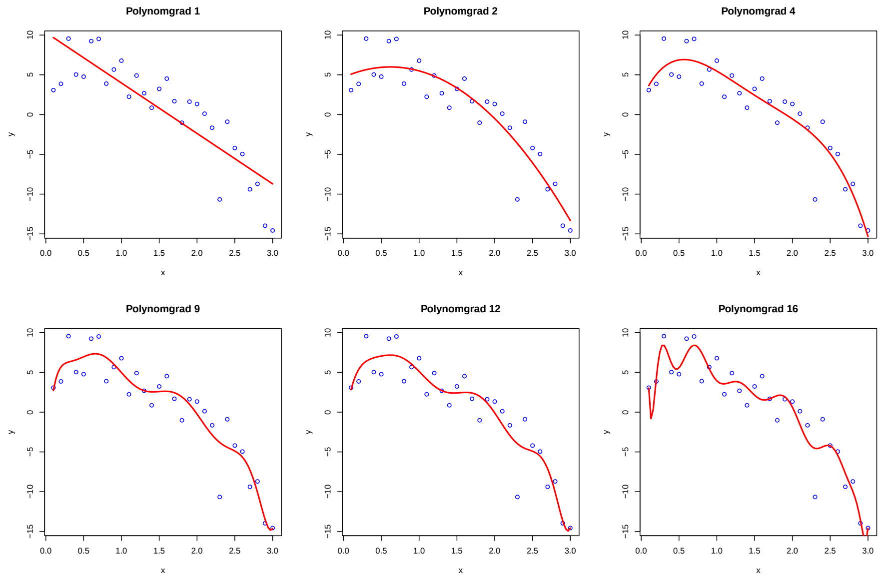


Gelenkfunktion $\phi(y) = \boldsymbol{x}^\top\boldsymbol{a}$

Generalized Linear Model

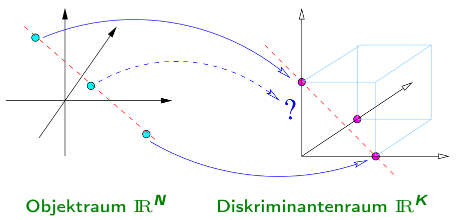
Überanpassungseffekt bei Ausgleichpolynomen

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



Maskierungseffekt

Lineare Quadratmitteldiskriminanten in Mehrklassensituationen



Kollineare Klassenzentren

$\boldsymbol{\mu}_\kappa = \beta_\kappa \boldsymbol{a} + \boldsymbol{b}$

$\rightsquigarrow$ kollineare Diskriminatenvektoren

$\boldsymbol{h}(\boldsymbol{\mu}_\kappa) = \beta_\kappa \tilde{\boldsymbol{a}} + \tilde{\boldsymbol{b}}$

(ideal $=$ κ-te Einheitsvektoren)

Quadratmittel

Minimiere den quadratischen Vorhersagefehler

$$\sum_t (y_t - \boldsymbol{a}^\top \boldsymbol{x}_t)^2$$

⊖ negative Werte!

⊖ nicht normiert!

Logit

Maximiere Datenwahrsch'keit

$$\prod_t \mathrm{P}(y_t|\boldsymbol{x}_t)$$

Probit

Maximiere Wahrsch'keitssumme

$$\sum_t \mathrm{P}(y_t|\boldsymbol{x}_t)$$

mit Posterior-Wahrscheinlichkeiten der logistischen Form $\mathrm{P}(\Omega_1|\boldsymbol{x}) \propto e^{\boldsymbol{a}^\top \boldsymbol{x}}$

Lineare logistische Regression

Zweiklassenmodell

Lineares Vorhersagemodell für die **log-odds**

$$\log \frac{\mathrm{P}(\Omega_1 \mid \boldsymbol{x})}{\mathrm{P}(\Omega_0 \mid \boldsymbol{x})} \stackrel{!}{=} h(\boldsymbol{x}) \;=\; \boldsymbol{a}^\top \boldsymbol{x}$$

Mehrklassenmodell

$K-1$ Modelle für logarithmierte Kontrastwahrscheinlichkeiten

$$\log \frac{\mathrm{P}(\Omega_\lambda \mid \boldsymbol{x})}{\mathrm{P}(\Omega_K \mid \boldsymbol{x})} \stackrel{!}{=} h_\lambda(\boldsymbol{x}) \;=\; \boldsymbol{a}_\lambda^\top \boldsymbol{x}$$

für alle $1 \le \lambda < K$.

Konsistente W'keiten

Alle $\mathrm{P}(\Omega_\lambda|\boldsymbol{x}) \in [0,1]$.
Alle Odds $\in [0,+\infty]$.
Log-odds $\in [-\infty,+\infty]$.

Umkehrformeln

Alle Klassen $\lambda \neq K$

$$p_\lambda(\boldsymbol{x}) \;=\; \frac{\exp(\boldsymbol{a}_\lambda^\top \boldsymbol{x})}{1 + \sum_{\kappa \neq K} \exp(\boldsymbol{a}_\kappa^\top \boldsymbol{x})}$$

Referenzklasse $\lambda = K$

$$p_\lambda(\boldsymbol{x}) \;=\; \frac{1}{1 + \sum_{\kappa \neq K} \exp(\boldsymbol{a}_\kappa^\top \boldsymbol{x})}$$

bzw. $p_1(\boldsymbol{x}) =$
$1 - p_0(\boldsymbol{x}) = \frac{e^{\boldsymbol{a}^\top \boldsymbol{x}}}{(1+e^{\boldsymbol{a}^\top \boldsymbol{x}})}$.

Maximum-Likelihood-Schätzung

Vereinfachter Fall: $K = 2$

Lemma

Für das binäre logistische Modell $p_1(\boldsymbol{x}) \propto \exp(\boldsymbol{a}^\top \boldsymbol{x})$ *mit den Lerndaten* $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^N \times \{1,0\}$, $t = 1,\ldots,T$ *gilt:*

1. *Die ML-Zielgröße besitzt die Darstellung*

$$\ell(\boldsymbol{a}) \;=\; \log \prod_{t=1}^{T} p_{y_t}(\boldsymbol{x}_t) \;=\; \sum_{t=1}^{T} \Big\{ y_t \cdot \boldsymbol{a}^\top \boldsymbol{x}_t \;-\; \log\big[1 + \exp(\boldsymbol{a}^\top \boldsymbol{x}_t)\big] \Big\} \;.$$

2. *Für ihren Gradientenvektor der partiellen Ableitungen gilt:*

$$\nabla_{\boldsymbol{a}} \;=\; \frac{\partial \ell(\boldsymbol{a})}{\partial \boldsymbol{a}} \;=\; \sum_{t=1}^{T} \boldsymbol{x}_t \cdot (y_t - p_1(\boldsymbol{x}_t)) \;=\; \boldsymbol{X}^\top \cdot (\boldsymbol{y} - \boldsymbol{p})$$

3. *Für ihre Hessematrix der gemischten partiellen Ableitungen gilt*

$$\boldsymbol{H}_{\boldsymbol{a}} \;=\; \frac{\partial^2 \ell(\boldsymbol{a})}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} \;=\; -\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top \cdot p_1(\boldsymbol{x}_t) \cdot (1 - p_1(\boldsymbol{x}_t)) \;=\; -\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \;,$$

wobei $\boldsymbol{W} = \mathrm{diag}(w_1,\ldots,w_T)$ *und* $w_t = p_1(\boldsymbol{x}_t) \cdot (1 - p_1(\boldsymbol{x}_t))$ *bezeichne.*

Der IRLS-Algorithmus

„Iteratively Reweighted Least Squares"

(Algorithmus)

1. INITIALISIERUNG $\boldsymbol{a} \leftarrow \boldsymbol{0}$
2. NEWTON-RAPHSON-SCHRITT

$$\boldsymbol{a} \;\leftarrow\; \boldsymbol{a} + \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}^\top \cdot (\boldsymbol{y} - \boldsymbol{p})$$

   Die diagonale Skalierungsmatrix $\boldsymbol{W} \in \mathbb{R}^{T \times T}$ hat Einträge $w_{tt} = p_t \cdot (1 - p_t)$, $p_t = \hat{p}_1(\boldsymbol{x}_t)$.

3. TERMINIERUNG
   Prüfe Abbruchbedingung; gehe $\rightsquigarrow$ 2 oder ENDE.

Newton-Raphson-Optimierungsschritt

Gradientenaufstieg mit quadratisch berechneter Schrittweite

$$\boldsymbol{a} \;\leftarrow\; \boldsymbol{a} - \boldsymbol{H}_{\boldsymbol{a}}^{-1} \cdot \nabla_{\boldsymbol{a}}$$

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression **Logitmodell** Präferenzen CART Σ

IRLS und Regularisierung

Was heißt eigentlich *„wiederholte Neugewichtung"*?

Newtonschritt = gewichtete lineare Regression

$$\begin{aligned}\boldsymbol{a}^* \;&\leftarrow\; \boldsymbol{a} + \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}^\top \cdot (\boldsymbol{y} - \boldsymbol{p}) \\ &= (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^\top \boldsymbol{W}^{1/2} \cdot \boldsymbol{W}^{1/2} \cdot \underbrace{(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p}))}_{\boldsymbol{z}}\end{aligned}$$

$\boldsymbol{a}^*$ ist Lösung der **gewichteten** Regressionsaufgabe $\left\|\boldsymbol{W}^{1/2} \cdot (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{a})\right\|^2 \overset{!}{\to}$ MIN

WLS-Regularisierung in jedem Newtonschritt

Löse gewichtetes GNG-System $\boldsymbol{R}_W \boldsymbol{a} = \boldsymbol{m}_W$ mit $\boldsymbol{R}_W = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ und $\boldsymbol{m}_W = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z}$ mittels regularisierter Koeffizientenmatrix:

$$\boldsymbol{R}_{W,\lambda} \overset{\text{def}}{=} (\boldsymbol{R}_W)_\lambda = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + \lambda \cdot \boldsymbol{E}$$

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression **Logitmodell** Präferenzen CART Σ

Maximum-Likelihood-Schätzung

Allgemeiner Fall: $K \geq 2$

Lemma

Für das logistische Modell $p_\lambda(\boldsymbol{x}) \propto \exp(\boldsymbol{a}_\lambda^\top \boldsymbol{x})$ mit den Lerndaten $(\boldsymbol{x}_t, g_t) \in \mathbb{R}^N \times \{1, \ldots, K\}$, $t = 1, \ldots, T$ gilt:

1. *Die ML-Zielgröße besitzt die Darstellung*

$$\ell(\boldsymbol{A}) = \log \prod_{t=1}^{T} p_{g_t}(\boldsymbol{x}_t) = \sum_{t=1}^{T} \left\{ \boldsymbol{a}_{g_t}^\top \boldsymbol{x}_t - \log \sum_\nu e^{\boldsymbol{a}_\nu^\top \boldsymbol{x}_t} \right\}, \quad \boldsymbol{a}_K = \boldsymbol{0} \in \mathbb{R}^N .$$

2. *Für die $K \cdot N$ partiellen Ableitungen ihrer Gradientenmatrix gilt:*

$$\frac{\partial \ell(\boldsymbol{A})}{\partial a_{\lambda,i}} = \sum_{t=1}^{T} x_{t,i} \cdot (y_{t,\lambda} - p_\lambda(\boldsymbol{x}_t))$$

3. *Für die $K^2 \cdot N^2$ gemischten partiellen Ableitungen ihres Hessetensors gilt:*

$$\frac{\partial \ell^2(\boldsymbol{A})}{\partial a_{\lambda,i} \cdot \partial a_{\kappa,j}} = \sum_{t=1}^{T} x_{t,i} x_{t,j} \cdot (p_\lambda(\boldsymbol{x}_t) \cdot p_\kappa(\boldsymbol{x}_t) - \delta_{\lambda,\kappa} \cdot p_\lambda(\boldsymbol{x}_t))$$

Es bezeichne $y_{t,\lambda} = \delta_{g_t,\lambda}$ die Klassenindikatorfunktion der Lerndaten.

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression **Logitmodell** Präferenzen CART Σ

Das Probit-Modell $(K = 2)$

Logistisches Wahrscheinlichkeitsmodell

mit einer **additiven** Zielfunktion

$$p_y(\boldsymbol{x}) = \frac{e^{y \cdot \boldsymbol{a}^\top \boldsymbol{x}}}{1 + e^{\boldsymbol{a}^\top \boldsymbol{x}}}, \qquad \ell(\boldsymbol{a}) = \sum_{t=1}^{T} p_{y_t}(\boldsymbol{x}_t) \overset{!}{\to} \text{MAX}$$

Gradientenvektor

$$\nabla_{\boldsymbol{a}} \ell(\boldsymbol{a}) = \sum_{t=1}^{T} p_{y_t}(\boldsymbol{x}_t) \cdot \{y_t - p_1(\boldsymbol{x}_t)\} \cdot \boldsymbol{x}_t = \boldsymbol{X}^\top \boldsymbol{Q}(\boldsymbol{y} - \boldsymbol{p})$$

mit $\boldsymbol{Q} = \mathrm{diag}(\{p_{y_t}(\boldsymbol{x}_t)\}_t)$

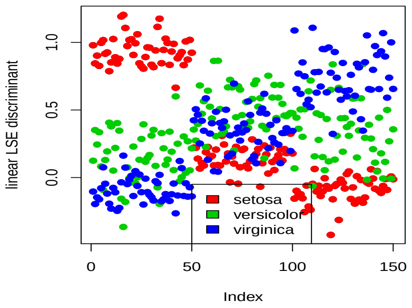
Hessematrix

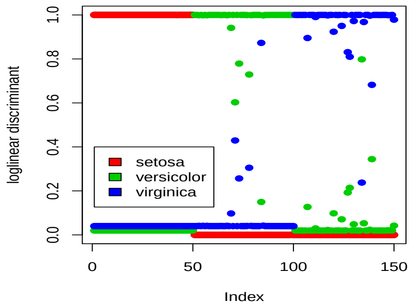
$$\boldsymbol{H}_{\boldsymbol{a}} = \sum_{t=1}^{T} p_{y_t}(\boldsymbol{x}_t) \cdot \{(y_t - p_1(\boldsymbol{x}_t))^2 + p_1^2(\boldsymbol{x}_t) - p_1(\boldsymbol{x}_t)\} \cdot \boldsymbol{x}_t \boldsymbol{x}_t^\top = -\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{P} \boldsymbol{X}$$

mit $\boldsymbol{W} = \mathrm{diag}(\{p_1(\boldsymbol{x}_t) \cdot p_0(\boldsymbol{x}_t)\}_t)$ und $\boldsymbol{P} = \mathrm{diag}(\{p_{y_t}(\boldsymbol{x}_t) - p_{1-y_t}(\boldsymbol{x}_t)\}_t)$

Reklassifikationsexperiment — Irisblüten-Datensatz

3 Klassen · 4 numerische Attribute · 50+50+50 Objekte





Quadratmittelmodell

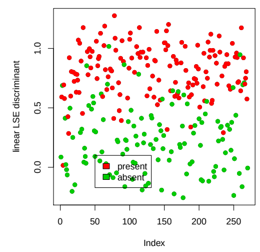
3 affine Prädiktoren $\boldsymbol{a}_\lambda \in \mathbb{R}^5$
Starke Schwankung um 1 und 0
Vertauschungen $\left\{ \begin{matrix} \texttt{'versicolor'} \\ \texttt{'virginica'} \end{matrix} \right\}$

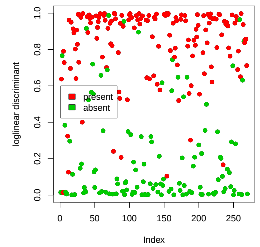
Loglinearmodell

2 affine Prädiktoren $\boldsymbol{a}_\lambda \in \mathbb{R}^5$
Fast alle Wahrsch'keiten bei $\{0, 1\}$
Fast perfekte Klassenidentifikation

Reklassifikationsexperiment — Herzkrankheiten-Datensatz

2 Klassen · 13 diskrete & numerische Attribute · 270 Objekte





Auszug Datenfriedhof

70.0 1.0 4.0 130.0 322.0 0.0 2.0 109.0 0.0 2.4 2.0 3.0 3.0 2
67.0 0.0 3.0 115.0 564.0 0.0 2.0 160.0 0.0 1.6 2.0 0.0 7.0 1
57.0 1.0 2.0 124.0 261.0 0.0 0.0 141.0 0.0 0.3 1.0 0.0 7.0 2
64.0 1.0 4.0 128.0 263.0 0.0 0.0 105.0 1.0 0.2 2.0 1.0 7.0 1
74.0 0.0 2.0 120.0 269.0 0.0 2.0 121.0 1.0 0.2 1.0 1.0 3.0 1
65.0 1.0 4.0 120.0 177.0 0.0 0.0 140.0 0.0 0.4 1.0 0.0 7.0 1
56.0 1.0 3.0 130.0 256.0 1.0 2.0 142.0 1.0 0.6 2.0 1.0 6.0 2
... ...

Attribute, Skalen, Werte

1.	age	$(\mathbb{R})$	-0.02511018
2.	sex	$\{male, female\}$	1.89901910
3.	chest pain	$\{A, B, C, D\}$	1.741, 0.784, 2.748
4.	blood pressure	$(\mathbb{R})$	0.03110868
5.	serum cholestoral	$(\mathbb{R})$	0.00655756
6.	fasting blood sugar	$\{T, F\}$	-0.37604461
	... ...		... ...
13	thal	$\{normal, fixed, defect\}$	-0.318, 1.468
	intercept		-7.68704469

Dualisierung der Regressionsaufgabe

Der Schlüssel zum Kerneltrick für Quadratmittel- & logistische Prädiktoren

Definition

Für eine Datenmatrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^\top$ des $\mathbb{R}^N$ bezeichne $\mathrm{Lin}(\boldsymbol{X})$ die **lineare Hülle** der Vektoren und $\mathrm{Lin}(\boldsymbol{X}^\perp)$ ihren **Orthogonalraum**.

Lineare Hülle

Die Menge aller Linearkombinationen der Matrixzeilen $\boldsymbol{x}_t$; sie bildet den kleinsten Untervektorraum von $\mathbb{R}^N$, der alle $\boldsymbol{x}_t$ enthält.

$$\mathrm{Lin}(\boldsymbol{X}) \;=\; \{\boldsymbol{X}^\top \boldsymbol{a} \mid \boldsymbol{a} \in \mathbb{R}^T\}$$

Orthogonalraum

Die Menge aller Vektoren, die auf allen Matrixzeilen $\boldsymbol{x}_t$ senkrecht stehen; sie bildet den größten Untervektorraum von $\mathbb{R}^N$, der keines der $\boldsymbol{x}_t$ enthält.

$$\mathrm{Lin}(\boldsymbol{X}^\perp) \;=\; \{\boldsymbol{z} \in \mathbb{R}^N \mid \boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}\}$$

Lemma

Lineare Hülle und Orthogonalraum spannen stets den Gesamtraum auf:

$$\mathbb{R}^N \;=\; \mathrm{Lin}(\boldsymbol{X}) \;\oplus\; \mathrm{Lin}(\boldsymbol{X}^\perp)$$

Darstellungssatz für QM-Lösungen

Endlichdimensionaler Spezialfall des Satzes von Kimeldorf & Wahba (1971)

Satz

Die regularisierten (unregularisierten) und gewichteten (ungewichteten) Quadratmittelaufgaben mit den Normalengleichungen

$$\begin{aligned} \boldsymbol{R} \cdot \boldsymbol{a} &= \boldsymbol{m} && \text{(LSE)} \\ \boldsymbol{R}_\lambda \cdot \boldsymbol{a} &= \boldsymbol{m} && \text{(RLSE)} \\ \boldsymbol{R}_w \cdot \boldsymbol{a} &= \boldsymbol{m}_w && \text{(WLSE)} \\ \boldsymbol{R}_{w,\lambda} \cdot \boldsymbol{a} &= \boldsymbol{m}_w && \text{(RWLSE)} \end{aligned}$$

besitzen jeweils mindestens eine Lösung, die sogar als Linearkombination der Datenvektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ darstellbar ist, d.h. es gilt:

$$\boldsymbol{a}^* \in \mathrm{Lin}(\boldsymbol{X})$$

Bezeichnungen

für die nicht normierten und unzentrierten Momente:

$$\begin{aligned} \boldsymbol{m} &= \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{R} &= \boldsymbol{X}^\top \boldsymbol{X} \\ \boldsymbol{R}_\lambda &= \boldsymbol{R} + \lambda \boldsymbol{E} \\ \boldsymbol{m}_w &= \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z} \\ \boldsymbol{R}_w &= \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \\ \boldsymbol{R}_{w,\lambda} &= \boldsymbol{R}_w + \lambda \boldsymbol{E} \end{aligned}$$

Beweis.

- REPRÄSENTATION FÜR LSE-LÖSUNG
  Ist $\boldsymbol{a} = \boldsymbol{a}_0 + \boldsymbol{a}_\perp$ mit $\boldsymbol{a}_0 \in \operatorname{Lin}(\boldsymbol{X})$ und $\boldsymbol{a}_\perp \in \operatorname{Lin}(\boldsymbol{X}^\perp)$ eine Lösung der GNG $\boldsymbol{Ra} = \boldsymbol{m}$, so gilt:

$$\boldsymbol{m} \;=\; \boldsymbol{Ra} \;=\; \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_0 + \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_\perp \;=\; \boldsymbol{R}\boldsymbol{a}_0$$

  Wir können folglich auch eine Lösung in $\operatorname{Lin}(\boldsymbol{X})$ finden.
- REPRÄSENTATION FÜR RLSE-LÖSUNG
  Ist $\boldsymbol{a} = \boldsymbol{a}_0 + \boldsymbol{a}_\perp$ eine Lösung der GNG $\boldsymbol{R}_\lambda \boldsymbol{a} = \boldsymbol{m}$, so gilt:

$$\boldsymbol{m} \;=\; \boldsymbol{R}_\lambda \boldsymbol{a} \;=\; \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a} + \lambda \boldsymbol{a}_0 + \lambda \boldsymbol{a}_\perp$$

  Da sowohl $\boldsymbol{m} = \boldsymbol{X}^\top \boldsymbol{y}$ als auch $\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}$ und $\lambda \boldsymbol{a}_0$ offensichtlich aus $\operatorname{Lin}(\boldsymbol{X})$ sind, ist das auch für den verbleibenden Ausdruck $\lambda \boldsymbol{a}_\perp$ der Fall. Wegen $\lambda > 0$ folgt $\boldsymbol{a}_\perp = \boldsymbol{0}$, also ist $\boldsymbol{a} = \boldsymbol{a}_0$ zwingend aus der linearen Hülle von $\boldsymbol{X}$.
- REPRÄSENTATION FÜR WLSE-LÖSUNG
  Im IRLS-Schritt sei $\boldsymbol{a} = \boldsymbol{a}_0 + \boldsymbol{a}_\perp$ eine Lösung der GNG $\boldsymbol{R}_w \boldsymbol{a} = \boldsymbol{m}_w$. Dann gilt:

$$\boldsymbol{m}_w \;=\; \boldsymbol{R}_w \boldsymbol{a} \;=\; \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \cdot \boldsymbol{a}_0 + \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \cdot \boldsymbol{a}_\perp \;=\; \boldsymbol{R}_w \boldsymbol{a}_0$$

- REPRÄSENTATION FÜR RWLSE-LÖSUNG
  Für die Lösung $\boldsymbol{a} = \boldsymbol{a}_0 + \boldsymbol{a}_\perp$ im regularisierten IRLS-Schritt gilt wie bei RLSE:

$$\boldsymbol{m}_w \;=\; \boldsymbol{R}_{w,\lambda} \boldsymbol{a} \;=\; \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + \lambda \boldsymbol{E}\right) \cdot (\boldsymbol{a}_0 + \boldsymbol{a}_\perp) \;=\; \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \boldsymbol{a} + \lambda \boldsymbol{a}_0 + \lambda \boldsymbol{a}_\perp$$

Regularisierung dualisierter QM-Aufgaben

Ungewichteter und gewichteter Fall

Lemma

*Die Lösungen der dualisierten **LSE-Aufgabe** lauten je nach Regularisierungstechnik:*

$$\begin{aligned}
\boldsymbol{b}^* &= \boldsymbol{G}^{-1} \cdot \boldsymbol{y} & \varepsilon(\boldsymbol{b}) &= \|\boldsymbol{y} - \boldsymbol{Gb}\|^2 \\
\boldsymbol{b}^* &= (\boldsymbol{G} + \lambda \boldsymbol{E})^{-1} \cdot \boldsymbol{y} & \varepsilon_\lambda(\boldsymbol{b}) &= \|\boldsymbol{y} - \boldsymbol{Gb}\|^2 + \lambda \cdot \left\|\boldsymbol{X}^\top \boldsymbol{b}\right\|^2 \\
\boldsymbol{b}^* &= (\boldsymbol{G}^2 + \lambda \boldsymbol{E})^{-1} \cdot \boldsymbol{Gy} & \varepsilon'_\lambda(\boldsymbol{b}) &= \|\boldsymbol{y} - \boldsymbol{Gb}\|^2 + \lambda \cdot \|\boldsymbol{b}\|^2
\end{aligned}$$

Lemma

*Die Lösungen der dualisierten **WLSE-Aufgabe** lauten je nach Regularisierungstechnik:*

$$\begin{aligned}
\boldsymbol{b}^* &= \boldsymbol{G}^{-1} \cdot \boldsymbol{z} & \varepsilon_W(\boldsymbol{b}) &= \|\boldsymbol{z} - \boldsymbol{Gb}\|_W^2 \\
\boldsymbol{b}^* &= (\boldsymbol{WG} + \lambda \boldsymbol{E})^{-1} \cdot \boldsymbol{Wz} & \varepsilon_{W,\lambda}(\boldsymbol{b}) &= \|\boldsymbol{z} - \boldsymbol{Gb}\|_W^2 + \lambda \cdot \left\|\boldsymbol{X}^\top \boldsymbol{b}\right\|^2 \\
\boldsymbol{b}^* &= (\boldsymbol{GWG} + \lambda \boldsymbol{E})^{-1} \cdot \boldsymbol{GWz} & \varepsilon'_{W,\lambda}(\boldsymbol{b}) &= \|\boldsymbol{z} - \boldsymbol{Gb}\|_W^2 + \lambda \cdot \|\boldsymbol{b}\|^2
\end{aligned}$$

$\mathrm{MSE}^\perp$ — die dualisierte Quadratmittelaufgabe

Speicheraufwand $\mathrm{O}(T^2)$ und Rechenaufwand $\mathrm{O}(T^3)$

Duale Lösungsdarstellung

als Linearkombination der Objektvektoren:

$$\boldsymbol{a} \;=\; \boldsymbol{X}^\top \boldsymbol{b} \;=\; \sum_{t=1}^{T} b_t \cdot \boldsymbol{x}_t \;, \qquad \boldsymbol{b} \in \mathbb{R}^T$$

Duale Regressionsfehlerformel

in Abhängigkeit vom Vektor $\boldsymbol{b}$ der Lösungskoeffizienten:

$$\varepsilon(\boldsymbol{b}) \;=\; \left\|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{X}^\top \boldsymbol{b}\right\|^2 \;\overset{!}{\to}\; \text{MIN}$$

Duale Gauß'sche Normalengleichungen

Lineares Gleichungssystem (Dimension $T \times T$) mit Gram'scher Matrix:

$$\boldsymbol{G}^2 \cdot \boldsymbol{b} \;=\; \boldsymbol{G} \cdot \boldsymbol{y} \;, \qquad \boldsymbol{G} \;=\; \boldsymbol{X} \cdot \boldsymbol{X}^\top$$

Beweis.

- UNREGULARISIERTE LÖSUNG:
  Der Gradientenvektor der Zielgröße

$$\varepsilon(\boldsymbol{b}) \;=\; \|\boldsymbol{y} - \boldsymbol{Gb}\|^2 \;=\; \boldsymbol{y}^\top \boldsymbol{y} - 2 \cdot \boldsymbol{b}^\top \boldsymbol{Gy} + \boldsymbol{b}^\top \boldsymbol{G}^2 \boldsymbol{b}$$

  lautet

$$\nabla_{\boldsymbol{b}} \varepsilon(\boldsymbol{b}) \;=\; \boldsymbol{0} - 2 \cdot \boldsymbol{Gy} + 2 \cdot \boldsymbol{G}^2 \boldsymbol{b} \,.$$

  Nullsetzen ergibt die GNG. Unter der Annahme einer regulären Gramschen Matrix ergibt sich die Lösung durch Multiplikation beider Gleichungsseiten mit $\boldsymbol{G}^{-2}$.
- REGULARISIERTE LÖSUNG I:
  Wir regularisieren im Vektorraum $\mathbb{R}^N$; der Fehlerterm besitzt den Gradientenvektor

$$\nabla_{\boldsymbol{b}} \varepsilon_\lambda(\boldsymbol{b}) \;=\; -2\boldsymbol{Gy} + 2\boldsymbol{G}^2 \boldsymbol{b} + 2\lambda \cdot \boldsymbol{Gb} \;=\; -2\boldsymbol{G} \cdot (\boldsymbol{y} - (\boldsymbol{G} + \lambda \boldsymbol{E}) \cdot \boldsymbol{b})$$

  Da $\boldsymbol{G}_\lambda$ regulär ist für $\lambda > 0$ liefert $\boldsymbol{b} = \boldsymbol{G}_\lambda^{-1} \boldsymbol{y}$ eine Lösung.
- REGULARISIERTE LÖSUNG II:
  Wir regularisieren im Vektorraum $\mathbb{R}^T$; der Fehlerterm besitzt den Gradientenvektor

$$\nabla_{\boldsymbol{b}} \varepsilon_\lambda(\boldsymbol{b}) \;=\; -2\boldsymbol{Gy} + 2\boldsymbol{G}^2 \boldsymbol{b} + 2\lambda \cdot \boldsymbol{b} \;=\; -2 \cdot (\boldsymbol{Gy} - (\boldsymbol{G}^2 + \lambda \boldsymbol{E}) \cdot \boldsymbol{b})$$

  Da auch $(\boldsymbol{G}^2)_\lambda$ regulär ist für $\lambda > 0$ liefert $\boldsymbol{b} = (\boldsymbol{G}^2)_\lambda^{-1} \cdot \boldsymbol{Gy}$ eine Lösung.

$\square$

Beweis.
Das zweite Lemma dient der schrittweisen Berechnung und Regularisierung im IRLS-Algorithmus für loglineare Modelle.

An Stelle des Fehlerfunktionals $\|\boldsymbol{y} - \boldsymbol{Gb}\|^2$ wird

$$\|\boldsymbol{z} - \boldsymbol{Gb}\|_{\boldsymbol{W}}^2 \stackrel{\text{def}}{=} (\boldsymbol{z} - \boldsymbol{Gb})^\top \cdot \boldsymbol{W} \cdot (\boldsymbol{z} - \boldsymbol{Gb})$$

minimiert. Wir unterscheiden wieder zwischen der Regularisierung im Raum $\mathbb{R}^N$ und im Raum $\mathbb{R}^T$.

- Ist $\boldsymbol{G}$ invertierbar, so hängt die Lösung $\boldsymbol{b}^* = \boldsymbol{G}^{-1}\boldsymbol{b}$ nicht von der (diagonalen) Gewichtmatrix $\boldsymbol{W}$ ab, denn $\boldsymbol{z} \approx \boldsymbol{Gb}$ wird ja mit exakter Gleichheit erfüllt:

$$\nabla\varepsilon_{\boldsymbol{W}}(\boldsymbol{b}) = -2\boldsymbol{GWz} + 2\boldsymbol{GWGb} = -2\boldsymbol{GW} \cdot (\boldsymbol{z} - \boldsymbol{Gb})$$

- Bei Regularisierung im Raum $\mathbb{R}^N$ ergibt sich:

$$\nabla\varepsilon_{\boldsymbol{W},\lambda}(\boldsymbol{b}) = -2\boldsymbol{GWz} + 2\boldsymbol{GWGb} + 2\lambda\boldsymbol{Gb} = -2\boldsymbol{G} \cdot (\boldsymbol{Wz} - (\boldsymbol{WG})_\lambda \cdot \boldsymbol{b})$$

- Bei Regularisierung im Raum $\mathbb{R}^T$ ergibt sich:

$$\nabla\varepsilon'_{\boldsymbol{W},\lambda}(\boldsymbol{b}) = -2\boldsymbol{GWz} + 2\boldsymbol{GWGb} + 2\lambda\boldsymbol{b} = -2 \cdot (\boldsymbol{GWz} - (\boldsymbol{GWG})_\lambda \cdot \boldsymbol{b})$$

□

Kombinatorische Regression

Aufgabenstellung
Klassifikation von Texten $v \in \boldsymbol{\Omega} = \mathcal{V}^*$ über Wortschatz $\mathcal{V}$

Termexpansion
Binärattribute:
Wort-m-Tupel oder
Wort-m-Subsets

$$\phi : \boldsymbol{\Omega} \to \{0,1\}^{\mathcal{V}^m}$$

mit $\phi_{\boldsymbol{u}}(x) = \begin{cases} 1 & \boldsymbol{u} \in x \\ 0 & \boldsymbol{u} \notin x \end{cases}$

Loglinearmodell
der Dimension L^m bzw. $\binom{L}{m}$

Duales Loglinearmodell
Gramsche T^2-Matrix mit Einträgen

$$K(x_s, x_t) = \langle \phi(x_s), \phi(x_t) \rangle$$

Kombinat. Kernoperator

$$K(x,y) = \sum_{\boldsymbol{u} \in \mathcal{V}^m} \phi_{\boldsymbol{u}}(x) \cdot \phi_{\boldsymbol{u}}(y) = \begin{cases} |V_x^m \cap V_y^m| & \text{Tupel} \\ \binom{|V_x^1 \cap V_y^1|}{m} & \text{Subsets} \end{cases}$$

Die Zählaufgaben $|V_x^m \cap V_y^m|$ sind sehr effizient zu bewältigen.

Ordinale Regression

Reelle Quellattribute $\mathbb{X}_1, \ldots, \mathbb{X}_N$ ➪ geordnetes Zielattribut $\mathbb{Y} \in \{1, \ldots, L\}$

Nominales Attribut
A posteriori Verteilung

$$p_\ell(\boldsymbol{x}) \stackrel{\text{def}}{=} \mathrm{P}(\mathbb{Y} = \ell \mid \mathbb{X} = \boldsymbol{x})$$

Normierungsbedingung:

$$\sum_\ell p_\ell(\boldsymbol{x}) = 1$$

Ordinales Attribut
Kumulative a post. Verteilung

$$q_\ell(\boldsymbol{x}) \stackrel{\text{def}}{=} \mathrm{P}(\mathbb{Y} \le \ell \mid \mathbb{X} = \boldsymbol{x})$$

Skalenbindung:

$$\boldsymbol{x} \rightsquigarrow z(\boldsymbol{x}) \in J_\ell \subset \mathbb{R}$$

Nominale Beispiele
RedGreenBlue-Skala:
$\boldsymbol{p} = (1/2, 1/6, 1/3)$
Unfairer Würfel:
$\boldsymbol{p} = (0, 0, 0, 1/4, 1/4, 1/2)$

Ordinale Beispiele
HighMediumLow-Skala:
$\boldsymbol{p} = (1/2, 1/6, 1/3)$
Zensurenskala:
$\boldsymbol{p} = (1/2, 0, 0, 0, 1/2, 0)$

Müssen ordinale Verteilungen zwangsläufig „*unimodal*“ sein?

Postulat der verborgenen dichten Qualitätsskala

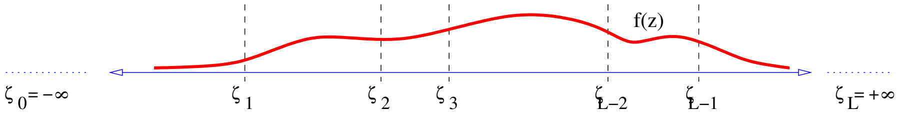
„Cumulative link model“ — Agresti 2002

Kumulatives Gelenkfunktionsmodell

Latente Variable $\mathbb{Z}$ auf der Skala $-\infty = \zeta_0 < \zeta_1 < \ldots < \zeta_L = +\infty$ mit

$$\mathbb{Y} = \ell \Leftrightarrow \mathbb{Z} \in (\zeta_{\ell-1}, \zeta_\ell] \quad \text{und} \quad \mathbb{Z} \sim f(\mu = h(\boldsymbol{x}), \sigma^2 = 1)$$

$h(\cdot)$ **Gelenkfunktion**, $f(\cdot)$ Verteilungsgesetz. $\left\{\begin{array}{c}\text{lognormal}\\ \text{normal}\\ \text{Extremwert}\\ \text{Cauchy}\end{array}\right\}$



Bemerkung

Es gilt $q_\ell(\boldsymbol{x}) = \mathrm{P}(\mathbb{Y} \le \ell \mid \mathbb{X} = \boldsymbol{x}) = \mathrm{P}(\mathbb{Z} \le \zeta_\ell \mid \mathbb{X} = \boldsymbol{x}) = F(\zeta_\ell - h(\boldsymbol{x}))$.

Proportional Odds Linear Regression

Lineares Binomialmodell für die Gelenkfunktion

POLR-Modell

Lineare Vorhersage der logarithmierten Chancenfunktionen:

$$\log \mathrm{odds}_\ell(\boldsymbol{x}) \stackrel{\text{def}}{=} \log \frac{q_\ell(\boldsymbol{x})}{1 - q_\ell(\boldsymbol{x})} = \log \frac{\mathrm{P}(\mathbb{Y} \le \ell \mid \mathbb{X} = \boldsymbol{x})}{\mathrm{P}(\mathbb{Y} > \ell \mid \mathbb{X} = \boldsymbol{x})} \stackrel{!}{=} \boldsymbol{a}^\top \boldsymbol{x} + \zeta_\ell$$

Bemerkungen

1. Normierung

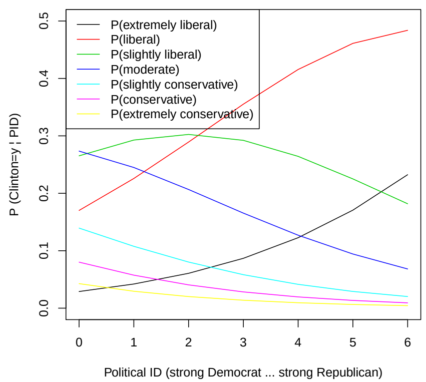
$$\sum_\ell p_\ell(x) = \sum_\ell (q_\ell(x) - q_{\ell-1}(x)) = q_L(x) - q_0(x) = 1 - 0$$

2. Monotonie $\quad k \le \ell \Rightarrow \zeta_k \le \zeta_\ell \Rightarrow \mathrm{loz}_k(x) \le \mathrm{loz}_\ell(x) \Rightarrow q_k(x) \le q_\ell(x)$
3. Proportionale Chancen

$$\log \frac{\mathrm{odds}_\ell(x)}{\mathrm{odds}_\ell(x')} = \boldsymbol{a}^\top (x - x') \qquad \text{ist unabhängig von } \ell$$

Beispiel — Präsidentschaftswahlen USA'96

http://www.stat.washington.edu/quinn/classes/536/data/nes96r.dat



Datensatz

944 Versuchspersonen
11 Attribute, u.a.:

- Pol/ID Clinton*
- Alter
  Bildungsgrad
  Einkommen
  Stimme für …
  TV-News/Woche
  Pol/ID selbst*
  Pol/ID Dole*

*) Pol/IDs in 7 Stufen

POLR-Datenanalyse

Zielattribut ≙ Einschätzung von Clintons politischer Haltung
Quellattribut ≙ politische Selbsteinschätzung
Fixiert: *3 TV/Woche, 44 Jahre, 12 Schuljahre, 35–40 Kilodollar*

Lernen von Präferenzrelationen

Objektive Präferenz

Aus einer Serie gewonnener, verlorener oder unentschiedener Partien $(\boldsymbol{x}_t, \boldsymbol{y}_t) \in \boldsymbol{\Omega} \times \boldsymbol{\Omega}$ ist eine passende Qualitätsrelation $(\boldsymbol{\Omega}, \preceq)$ zu lernen.

„Tourniermetapher“

Subjektive Präferenz

Aus einer Serie persönlicher Nennungen, Wertungen oder Reihungen $(\boldsymbol{s}_t, \boldsymbol{x}_t) \in \mathfrak{S} \times \boldsymbol{\Omega}$ ist eine **Schar** passender Qualitätsrelationen $(\boldsymbol{\Omega}, \preceq_s)_{s \in \mathfrak{S}}$ zu lernen.

„Jurorenmetapher“

Geschlossene Welten

Objektraum $\boldsymbol{\Omega}$ und/oder Subjektraum $\mathfrak{S}$ bilden ein endliches **Inventar**.
(Nominalattribut)

Offene Welten

Objekte u/o Subjekte sind durch ihre **Eigenschaften** charakterisiert.
(Attributvektoren)

Objektive Präferenz durch logistische Regression

Bilaterales Ereignismodell

$\mathbb{X} \hat{=}$ Objekt #1 (Herausforderer)
$\mathbb{Y} \hat{=}$ Objekt #2 (Gegner)
$\mathbb{Z} \hat{=}$ Resultat $\pm$„*Sieg*“ oder $\pm$„*Tor*“ …

Logistisch-lineares Erfolgsmodell

$$\log \text{odds}(\boldsymbol{x}, \boldsymbol{y}) \;=\; \underbrace{\boldsymbol{a}^\top \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{y} + \zeta}_{g(x) - h(y)}$$

Präferenzinterpretation

$\boldsymbol{x}$ hat immer dann bessere Gewinnchancen als $\boldsymbol{y}$ wenn $g(\boldsymbol{x}) > h(\boldsymbol{y})$ gilt.

Intervallordnung ?

Es gilt $p(\boldsymbol{x}, \boldsymbol{x}) \leq 1/2 \;\Leftrightarrow\; g_x \leq h_x$.

$$\boldsymbol{x} \succ \boldsymbol{y} \quad \text{gdw.} \quad [g_x, h_x] \sqsupset [g_y, h_y]$$

Fußballturnier

GER : BRA	3:1
USA : LBY	0:1
UK : IRAN	2:2

Punktestandbezogen

x_{GER}	x_{BRA}	$+$
x_{BRA}	x_{GER}	$-$
x_{USA}	x_{LBY}	$-$
x_{LBY}	x_{USA}	$+$
x_{UK}	x_{IRAN}	$-$
x_{IRAN}	x_{UK}	$-$

Torstandbezogen

x_{GER}	x_{BRA}	$+$	3
x_{GER}	x_{BRA}	$-$	87
x_{BRA}	x_{GER}	$+$	1
x_{BRA}	x_{GER}	$-$	89
…	…		
x_{UK}	x_{IRAN}	$+$	2
x_{UK}	x_{IRAN}	$-$	88
x_{IRAN}	x_{UK}	$+$	2
x_{IRAN}	x_{UK}	$-$	88

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Objektive Präferenz durch Proportional-Odds Regression

Trilaterales Ereignismodell

$\mathbb{X} \hat{=}$ Objekt #1 (Herausforderer)
$\mathbb{Y} \hat{=}$ Objekt #2 (Gegner)
$\mathbb{Z} \hat{=}$ Resultat aus $\{\xi_1, \xi_2, \xi_3\} = \{\gtrdot, \doteq, \lessdot\}$

POLR Erfolgsmodell

$$\log \text{odds}_\ell(\boldsymbol{x}, \boldsymbol{y}) \;=\; \underbrace{\boldsymbol{a}^\top \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{y} + \zeta_\ell}_{\boldsymbol{a}^\top(\boldsymbol{x}-\boldsymbol{y}) \,\pm\, \zeta}$$

Präferenzinterpretation

$\boldsymbol{x}$ hat immer dann bessere Gewinnchancen als $\boldsymbol{y}$ wenn $\log \text{odds}_1(\boldsymbol{x}, \boldsymbol{y}) > 0$ gilt, also

$$g_x := \boldsymbol{a}^\top \boldsymbol{x} \;>\; \boldsymbol{a}^\top \boldsymbol{y} + \zeta =: h_y$$

Semi-Ordnung !

$$\boldsymbol{x} \succ \boldsymbol{y} \quad \text{gdw.} \quad [g_x, g_x + \zeta] \sqsupset [g_y, g_y + \zeta]$$

Fußballturnier

GER : BRA	3:1
USA : LBY	0:1
UK : IRAN	2:2

Punktestandbezogen

x_{GER}	x_{BRA}	$\gtrdot$
x_{BRA}	x_{GER}	$\lessdot$
x_{USA}	x_{LBY}	$\lessdot$
x_{LBY}	x_{USA}	$\gtrdot$
x_{UK}	x_{IRAN}	$\doteq$
x_{IRAN}	x_{UK}	$\doteq$

Torstandbezogen

x_{GER}	x_{BRA}	$\gtrdot$	3
x_{GER}	x_{BRA}	$\doteq$	86
x_{GER}	x_{BRA}	$\lessdot$	1
x_{BRA}	x_{GER}	$\gtrdot$	1
x_{BRA}	x_{GER}	$\doteq$	86
x_{BRA}	x_{GER}	$\lessdot$	3
…	…		

Spezielle Form des POLR-Modells

$$\log \text{odds}_\ell(\boldsymbol{x}, \boldsymbol{y}) \;=\; \boldsymbol{a}^\top(\boldsymbol{x} - \boldsymbol{y}) + \begin{cases} -\infty & \ell = 0 \\ -\zeta & \ell = 1 \\ +\zeta & \ell = 2 \\ +\infty & \ell = 3 \end{cases}$$

Beweis.

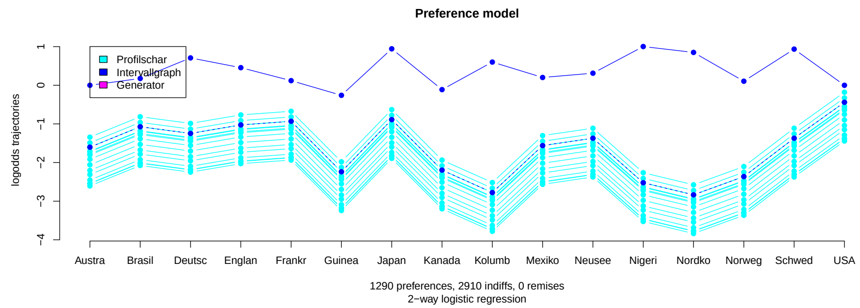
Aus der strukturellen Symmetrie
$\mathrm{P}(\gtrdot \mid \boldsymbol{x}, \boldsymbol{y}) = \mathrm{P}(\lessdot \mid \boldsymbol{y}, \boldsymbol{x})$ folgt:

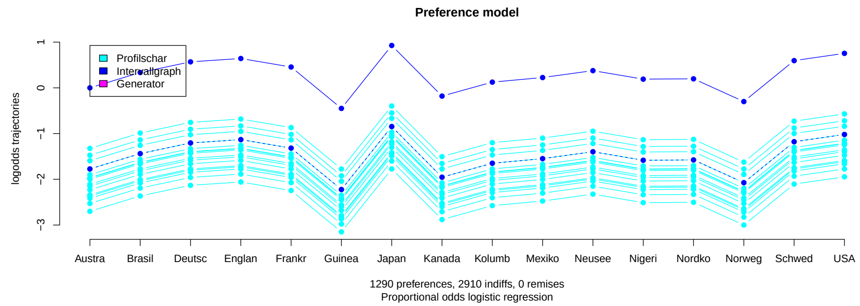
$$\begin{aligned}
\Rightarrow\quad & p_1(\boldsymbol{x}, \boldsymbol{y}) \;=\; p_3(\boldsymbol{y}, \boldsymbol{x}) \\
\Rightarrow\quad & q_1(\boldsymbol{x}, \boldsymbol{y}) - 0 \;=\; 1 - q_2(\boldsymbol{y}, \boldsymbol{x}) \\
\Rightarrow\quad & \frac{q_1(\boldsymbol{x}, \boldsymbol{y})}{1 - q_1(\boldsymbol{x}, \boldsymbol{y})} \;=\; \frac{1 - q_2(\boldsymbol{y}, \boldsymbol{x})}{q_2(\boldsymbol{y}, \boldsymbol{x})} \\
\Rightarrow\quad & \text{odds}_1(\boldsymbol{x}, \boldsymbol{y}) \;=\; \text{odds}_2^{-1}(\boldsymbol{y}, \boldsymbol{x}) \\
\Rightarrow\quad & +\log \text{odds}_1(\boldsymbol{x}, \boldsymbol{y}) \;=\; -\log \text{odds}_2(\boldsymbol{y}, \boldsymbol{x}) \\
\Rightarrow\quad & 0 \;=\; \boldsymbol{a}^\top \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{y} + \zeta_1 + \boldsymbol{a}^\top \boldsymbol{y} + \boldsymbol{b}^\top \boldsymbol{x} + \zeta_2 \\
\Rightarrow\quad & 0 \;=\; (\boldsymbol{a} + \boldsymbol{b})^\top (\boldsymbol{x} + \boldsymbol{y}) + (\zeta_1 + \zeta_2) \\
\Rightarrow\quad & \boldsymbol{b} = -\boldsymbol{a} \text{ und } \zeta_1 = -\zeta_2
\end{aligned}$$

□

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

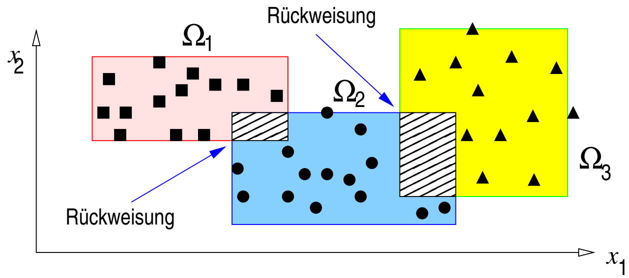
Beispiel — Frauenfußball-WM 2011





Parallelepiped-Klassifikator

Vollständige Konjunktion je zweier Literale $x_n \geq a_n$, $x_n \leq b_n$, $n = 1, \ldots, N$



Vorteile

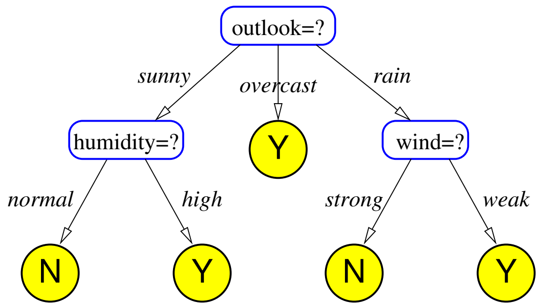
Extrem schnelle Lernphase
Effiziente Abrufphase
Klassengebiete intuitiv zu deuten
Nominalattribute handhabbar$^{\in}$

Nachteile

Achsenparallele Grenzen
Unimodale Klassengebiete
Ausgedehnte Rückweisungszonen
Keine Rückschlußwahrsch'keiten

Entscheidungsbaum

Hierarchie sequentieller Auswahlfragen („*multiple choice*“)



Sportwetterempfehlungen

Vier nominale Wetterlagevariablen gegeben

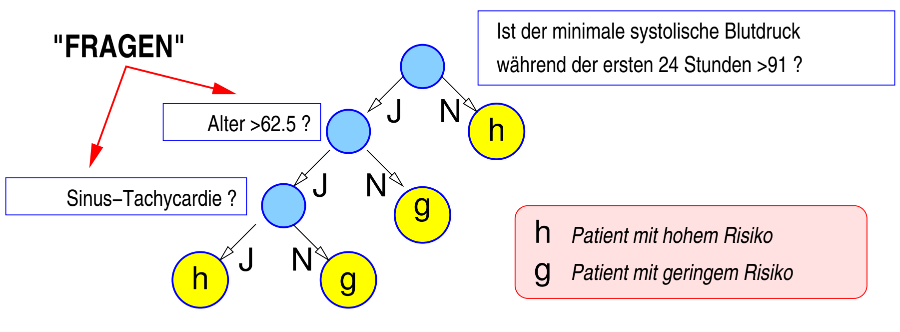
Fragetypus

Wertverzweigung

Klassifikationsziel:
„Ist dieses Wetter zum Tennisspielen geeignet?“

Binärer Entscheidungsbaum

Hierarchie sequentieller Ja/Nein-Fragen



Diagnose für Herzinfarktpatienten

19 Attribute gemessen bzw. erfragt
Patienten 30 Tage unter klinischer Beobachtung

Fragetypus

Schwellwertdichotomie
Zielwertdichotomie

Klassifikationsziel:
„Ist ein zweiter, diesmal tödlicher Infarkt eingetreten?“

Struktur eines Entscheidungsbaumes

Knoten ≙ Fragen

$\mathcal{B}$ bezeichnet die Menge aller Knoten.
Innere Knoten $\beta \in \mathcal{B}$ beherbergen eine **Entscheidungsfrage**:

$$Q(\beta): \ \mathbf{\Omega} \ \to \ \{1, \ldots, L\}$$

Kanten ≙ Antworten

Für $\beta \in \mathcal{B}$ ist $\beta^{\uparrow}$ der **Vorgängerknoten** und $\beta^{(1)}, \ldots, \beta^{(L)}$ sind die unmittelbaren **Nachfolger**.

Wurzelknoten ≙ Startposition

$\beta_{\triangle} \in \mathcal{B}$ besitzt keinen Vorgänger.
In $\beta_{\triangle}$ beginnt die Befragung des Objekts.

Blattknoten ≙ Ergebnis

Die $\beta \in \mathcal{B}_{\ell}$ besitzen keine Nachfolger, aber eine **Klassenmarkierung**:

$$\delta_{\ell} : \mathcal{B}_{\ell} \longrightarrow \{1, \ldots, K\}$$

Befragung der Attributwerte

Dichotomien („*Yin-Yang*“-Fragen) und Wertverzweigungen

Attribut-Wert-Gleichungen $\quad x_i \ = \ low$

bei nominalen Merkmalen
(das negative Literal $x_i \neq low$ ist dazu dual)

Attribut-Wert-Ungleichungen $\quad x_i \leq 3.14$

bei ordinalen Merkmalen
(auch $x_i \geq 17$ oder Intervalle $18 \leq x_i \leq 65$ denkbar)

Wertverzweigungen $\quad x_i \ = \ red|blue|green$

bei Attributen mit kleinem $|\mathcal{X}_n|$
(eine Nachfolgerkante je Attributwert)

Teilmengenzugehörigkeit $\quad x_i \ \in \ \{cloudy, rainy\}$

bei nominalen Attributen

Reguläre Ausdrücke $\quad x_i \ = \ ababb * c * ba$

bei Wort- oder Zeichenketten

Klassifikation eines Objekts

Hierarchisches Interview — „*Durchschleusen*“ bis zum Blattknoten

⟨Algorithmus⟩

1. INITIALISIERUNG
   Setze $\beta = \beta_{\triangle}(\mathcal{B})$.
2. BEFRAGUNG
   Reiche $\boldsymbol{x}$ gemäß $Q(\beta)$ an einen Kindknoten weiter:
   $$\beta \leftarrow \beta^{(i)}, \quad i = Q(\beta)(\boldsymbol{x})$$
3. TERMINIERUNG
   Ist β ein Blattknoten, so lautet das Resultat:
   $$\delta(\boldsymbol{x}) \ = \ \delta_{\ell}(\beta)$$
   Andernfalls $\rightsquigarrow$ 2.

⟨/Algorithmus⟩

Simultanes Schleusen einer Objektmenge

Definition

Ist $(\mathcal{B}, Q, \delta_{\ell})$ ein Entscheidungsbaum über $\mathbf{\Omega}$ und $\omega \subseteq \mathbf{\Omega}$ ein Datensatz, so definieren wir die **assoziierten Objektmengen** ω_{β} induktiv durch:

$$\omega(\beta) \ \stackrel{\text{def}}{=} \ \begin{cases} \omega & \beta = \beta_{\triangle} \\ \{\boldsymbol{x} \in \omega_{\beta} \mid Q(\beta)(\boldsymbol{x}) = j\} & \beta = \beta^{(j)} \end{cases}$$

Lemma

Ist $(\mathcal{B}, Q, \delta_{\ell})$ *ein Entscheidungsbaum über* $\mathbf{\Omega}$*, so gilt:*

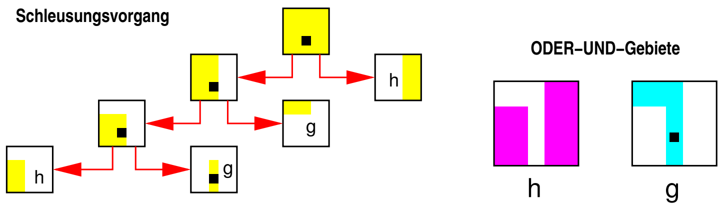
$$\mathbf{\Omega} \ = \ \biguplus_{\beta \in \mathcal{B}_{\ell}} \mathbf{\Omega}(\beta)$$

Der Entscheidungsbaum definiert ferner eine vollständige Zerlegung von $\mathbf{\Omega}$ *in Klassengebiete:*

$$\mathbf{\Omega} \ = \ \biguplus_{\kappa=1}^{K} \mathbf{\Omega}_{\kappa}, \qquad \mathbf{\Omega}_{\kappa} \ \stackrel{\text{def}}{=} \ \bigcup_{\delta_{\ell}(\beta)=\kappa} \mathbf{\Omega}(\beta)$$

Entscheidungsbäume als Hypothesen

Disjunktionen von Literalkonjunktionen



Beispiel

Intuitiv interpretierbare Klassenentscheidungen:

$$\begin{aligned} \boldsymbol{\Omega}_h &= (\{x_b \not\leq 91\} \wedge \{x_a \not\leq 62.5\} \wedge \{x_s \not\leq 0\}) \vee (\{x_b \leq 91\}) \\ \boldsymbol{\Omega}_g &= (\{x_b \not\leq 91\} \wedge \{x_a \not\leq 62.5\} \wedge \{x_s \leq 0\}) \vee (\{x_b \not\leq 91\} \wedge \{x_a \leq 62.5\}) \end{aligned}$$

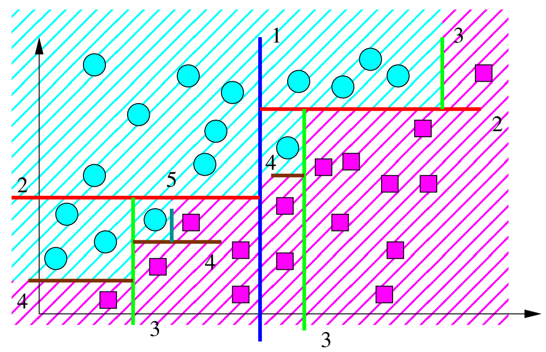
mit den Variablen (Merkmalen)

$$\begin{aligned} x_b &= \text{minimaler systolischer Blutdruck} \\ x_a &= \text{Alter des Patienten} \\ x_s &= \text{Sinus-Tachycardie? (0 oder 1)} \end{aligned}$$

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Entscheidungsbäume für numerische Attribute?

Rekursive Halbraumbildung nach sukzessiven Schwellwertabfragen $x_n \leq \theta$



Vom Entscheidungsbaum induzierte Klassengebiete

$$\hat{\Omega}_\kappa = \bigcup_{m=1}^{M_\kappa} \hat{\Omega}_{\kappa,m}\,, \qquad \hat{\Omega}_{\kappa,m} = \bigcap_{l=1}^{M_{\kappa,m}} H_{\kappa,m,l} = \text{Halbraum} \left\{ \begin{matrix} x_d \leq \theta \\ \text{oder} \\ x_d > \theta \end{matrix} \right\}$$

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Lernen eines Entscheidungsbaumes

aus klassenetikettierten Beispielobjekten: $\omega = \omega_1 \uplus \omega_2 \uplus \ldots \uplus \omega_K \subset \boldsymbol{\Omega}$

Trennschärfe

Der Baum soll die Beispiele möglichst *korrekt klassifizieren*.

➪ **Konsistenz**

Induktionskraft

Der Baum soll die Beispiele in geeigneter Weise *verallgemeinern*.

➪ **geringe Knotenzahl**

Hypothesenraum

Welche Größe? Welche Form?
Welches Attribut? Welche Frage?

➪ gigantische Auswahl an E-Bäumen

Lernen ≙ Suche

Vollständige Suche ist NP-hart.

➪ lokal optimierende Suche (Bergsteiger-Algorithmus, „*divide-and-conquer*“)

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

TDI-Lernalgorithmus

Gierige Top-Down Induktion von Entscheidungsbäumen

(Algorithmus)

1. INITIALISIERUNG
   Erzeuge einen Wurzelknoten $\beta = \beta_\triangle$ mit den assoziierten Stichproben $\omega_1, \ldots, \omega_K$.
2. STOPPTEST
   Ist β hinreichend **reinklassig**, so beende die lokale Konstruktion mit der Blattmarkierung
   $$\delta_\ell(\beta) = \underset{\kappa}{\operatorname{argmax}} |\omega_\kappa(\beta)|\,.$$
3. FRAGEAUSWAHL
   Wähle eine Frage $Q(\beta)$ mit maximaler Reduktion der **Entscheidungsunsicherheit**.
4. EXPANSION
   Bilde die Nachfolgerknoten $\beta^{(1)}, \ldots, \beta^{(L)}$ bezüglich $Q(\beta)$ und ihre assoziierten Stichproben
   $$\omega_\kappa(\beta^{(l)})\,, \quad l = 1, \ldots, L\,.$$
5. REKURSION
   Fahre mit den Nachfolgern $\beta^{(l)}$ von β bei Schritt 2 fort.

(Algorithmus)

Stoppkriterium

Wann endet der Züchtungsvorgang?

Lokale Stoppkriterien

Wann endet die Knotenexpansion in einem Blatt?

- Wenn $\omega(\beta)$ nur noch einen Datenvektor enthält.
- Wenn $\omega(\beta)$ nur noch Daten einer Klasse enthält. ➪ Konsistenz
- Wenn $|\omega(\beta)|$ eine gegebene Schranke unterschreitet.
- Wenn $|\omega(\beta)| - \max_\lambda |\omega_\lambda(\beta)|$ eine Schranke unterschreitet.

Überanpassung an die Lernbeispiele

Gefährlich in großen Bäumen durch *Zersplitterung* von ω auf die Blattknoten.

- *Ist es wirklich weise, einen konsistenten Baum zu konstruieren ?*

➪ globale a posteriori Stoppkriterien
a.k.a. Baumbeschneidungstechniken, „*pruning*“

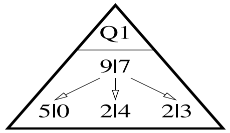
Die Frage nach der richtigen Frage

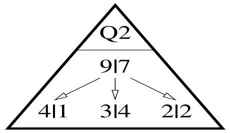
... bei Yuichiro Anzai im Autohaus ... (Beispiel)

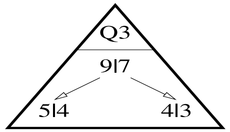
Japanische Gebrauchtfahrzeuge

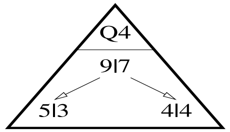
und ihre Veräußerungschancen am Markt

Objekt	cm^3	Türen	Autom.	Farbe	$x \overset{?}{\in} C$
x_1	2000	$2T$	*ja*	*hell*	$+$
x_2	2800	$4T$	*ja*	*hell*	$+$
x_3	2000	$2T$	*nein*	*dunkel*	$-$
x_4	1600	$4T$	*ja*	*dunkel*	$-$
x_5	1600	$4T$	*ja*	*hell*	$-$
x_6	2800	$4T$	*ja*	*dunkel*	$+$
x_7	2000	$4T$	*ja*	*hell*	$+$
x_8	2000	$5T$	*nein*	*hell*	$-$
x_9	1600	$2T$	*nein*	*hell*	$+$
x_{10}	2800	$5T$	*ja*	*hell*	$+$
x_{11}	2800	$5T$	*nein*	*dunkel*	$+$
x_{12}	2000	$4T$	*ja*	*dunkel*	$-$
x_{13}	1600	$2T$	*nein*	*dunkel*	$+$
x_{14}	2800	$2T$	*nein*	*dunkel*	$+$
x_{15}	1600	$4T$	*nein*	*hell*	$-$
x_{16}	2000	$5T$	*ja*	*dunkel*	$-$









Auswahlregel für die „*beste*“ nächste Frage

Vorgehensweise

Welches ist die (lokal) zielführendste Frage ?

1. Definiere Entscheidungsunsicherheit einer Häufigkeitsverteilung
2. Definiere Entscheidungsunsicherheit eines Baumknotens
3. Definiere Entscheidungsunsicherheit einer Frage (in β)
4. Definiere Entscheidungsunsicherheit eines Teilbaums (unter β)

Relative Klassenhäufigkeit

in der Teilstichprobe ω_β zum Knoten $\beta \in \mathcal{B}$:

$$\hat{p}_\kappa(\beta) = \frac{\text{Anzahl der } \Omega_\kappa\text{-Muster in } \beta}{\text{Anzahl aller Muster in } \beta} = \frac{|\omega_\kappa(\beta)|}{\sum_{\lambda=1}^{K} |\omega_\lambda(\beta)|}$$

➪ ML-Schätzwert für $\mathrm{P}(\boldsymbol{x} \in \Omega_\kappa \mid \boldsymbol{x} \in \boldsymbol{\Omega}_\beta)$

Entscheidungsunsicherheit

Gütemaß für den Entmischungsgrad einer Verteilung

Definition

Es sei $K \in \mathbb{N}$ und $\{p_\kappa \mid \kappa = 1, \ldots, K\}$ eine diskrete Wahrscheinlichkeitsverteilung. Eine Abbildung

$$\Im : \{p_1, \ldots, p_K\} \mapsto u \in \mathbb{R}$$

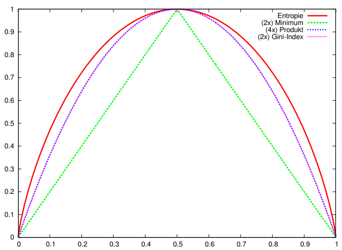
heißt Maß für die **Entscheidungsunsicherheit** (Homogenität, „*impurity*“), falls gilt:

1. Die Größe $\Im(\cdot)$ ist nichtnegativ.
2. $\Im(\cdot)$ ist maximal für die Gleichverteilung $p_\kappa \equiv 1/K$
3. $\Im(\cdot)$ ist minimal für die definiten Verteilungen

$$\boldsymbol{e}_\lambda = (\underbrace{0, \ldots, 0}_{\lambda-1}, 1, \underbrace{0, \ldots, 0}_{K-\lambda}), \quad \lambda \in \{1, \ldots, K\}$$

Homogenitätsmaße

Minimum · Produkt · Gini-Index · Entropie



Extremalwerte

	min	max
$\Im_m$	0	$^1/_K$
$\Im_p$	0	$^1/_{K^K}$
$\Im_g$	0	$1 - {}^1/_K$
$\Im_e$	0	$\log_2 K$

Lemma

Die folgenden Abbildungen sind (für festes $K \in \mathbb{N}$) Beispiele für Homogenitätsmaße:

$$\Im_m(\boldsymbol{p}) \stackrel{def}{=} \min_\kappa p_\kappa \qquad \Im_g(\boldsymbol{p}) \stackrel{def}{=} \sum_{\lambda \neq \kappa} p_\lambda \cdot p_\kappa$$

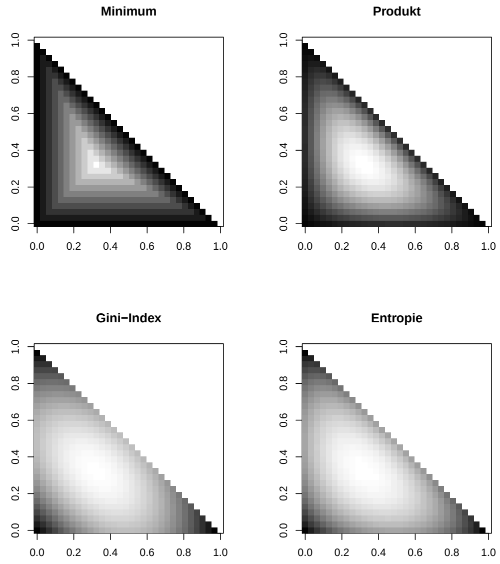
$$\Im_p(\boldsymbol{p}) \stackrel{def}{=} \prod_\kappa p_\kappa \qquad \Im_e(\boldsymbol{p}) \stackrel{def}{=} -\sum_\kappa p_\kappa \cdot \log_2 p_\kappa$$

Für den Gini-Index gilt $\Im_g(\boldsymbol{p}) = 1 - \|\boldsymbol{p}\|^2$.

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Homogenitätsmaße

Drei Ereignisse — Darstellung in der (p_1, p_2)-Ebene



Minimum/Produkt

Geringe Homogenität (Unsicherheit) wird bereits dann signalisiert, wenn nur eines der drei Ereignisse unwahrscheinlich ist.

➪ unbrauchbar

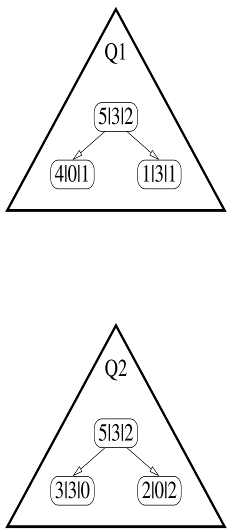
Entropie/Gini

Grundverschiedene Formeln, aber kaum unterschiedliche Funktionswerte.

➪ praktisch äquivalent

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Rechenbeispiel (Gini-Index)



Ausgangsknoten β

Der Knoten β beherbergt die Verteilung $\boldsymbol{p} = (0.5, 0.3, 0.2)$, also gilt

$$\Im_{\text{Gini}}(\boldsymbol{p}) = 1 - 0.25 - 0.09 - 0.04 = 0.62$$

Erste Frage Q_1

Die Entscheidungsunsicherheiten der Q_1-Nachfolger lauten

$$\Im_{\text{Gini}}(\beta^{(1)} \mid Q_1) = 1 - 0.64 - 0.04 = 0.32$$
$$\Im_{\text{Gini}}(\beta^{(2)} \mid Q_1) = 1 - 0.04 - 0.36 - 0.04 = 0.56$$

Die mittlere E.U. der Nachfolger bzw. ihre Reduktion betragen also

$$\Im_{\text{Gini}}(\beta \mid Q_1) = 0.5 \cdot 0.32 + 0.5 \cdot 0.56 = 0.44$$
$$\Delta_{Q_1} \Im_{\text{Gini}}(\beta) = 0.62 - 0.44 = 0.18$$

Zweite Frage Q_2

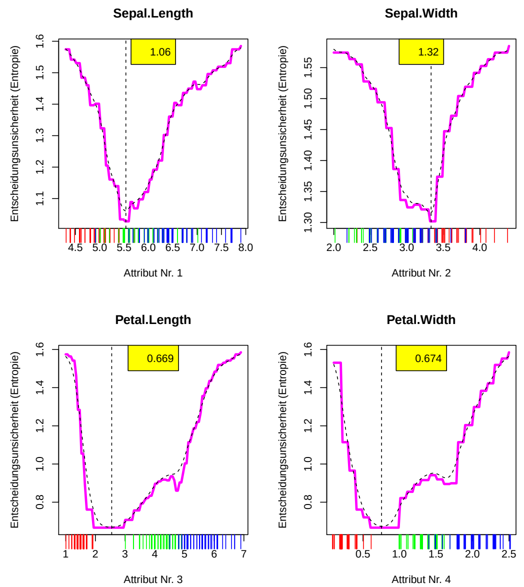
Auf dieselbe Weise errechnet sich für die konkurrierende Frage der Wert

$$\Delta_{Q_2} \Im_{\text{Gini}}(\beta) = 0.62 - 0.6 \cdot 0.5 - 0.4 \cdot 0.5 = 0.12$$

Folglich ist Q_1 der Frage Q_2 vorzuziehen.

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Rechenbeispiel (Entropiemaß)



IRIS-Datensatz

150 Objekte
4 Attribute
3 Kategorien $\begin{Bmatrix} 50 \\ 50 \\ 50 \end{Bmatrix}$

Wurzelknoten

Berechne für jedes Attribut x_n den EU-minimalen Schwellenwert θ_n

$$Q(\beta): \; x_3 \stackrel{?}{\leq} 2.65$$

Reduktion der Entscheidungsunsicherheit

Entscheidungsunsicherheit im Knoten β

$$\Im(\beta) \stackrel{\text{def}}{=} \Im(\hat{\boldsymbol{p}}^{(\beta)}) , \qquad \hat{p}_\kappa \stackrel{\text{def}}{=} \frac{|\omega_\kappa(\beta)|}{|\omega(\beta)|}$$

Verzweigungswahrscheinlichkeiten der Frage Q in β

$$\hat{\mathrm{P}}(\beta^{(i)}|\beta) \stackrel{\text{def}}{=} \frac{|\omega(\beta^{(i)})|}{|\omega(\beta)|} , \qquad i = 1, \ldots, L$$

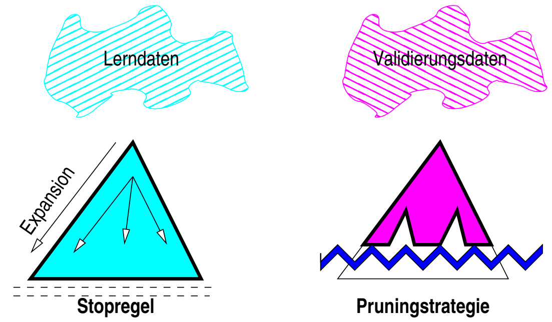
Entscheidungsunsicherheit nach der Frage Q in β

$$\Im(\beta|Q) \stackrel{\text{def}}{=} \sum_i \hat{\mathrm{P}}(\beta^{(i)}|\beta) \cdot \Im(\beta^{(i)})$$

Reduktion der Entscheidungsunsicherheit durch Q

$$\Delta_Q(\beta) \stackrel{\text{def}}{=} \Im(\beta) - \Im(\beta|Q)$$

Überanpassung an die Lerndaten



Fragmentierung der Lerndaten

- $\mathcal{B}$ reinklassig $\rightsquigarrow$ ω perfekt klassifiziert
- viele Lerndaten $\rightsquigarrow$ großer Entscheidungsbaum
- Insignifikante Fragen in unteren Zweigen
- Unzuverlässige Entscheidung in den Blättern
- Stoppregeln sind „*kurzsichtig*“

Abhilfe

- ⇩ „*early stopping*“
- ⇧ Züchten & Zurückschneiden

Aufspüren und Tilgen nutzloser Teilbäume

CART Pruning

Züchtung eines überangepaßten Baumes	Lerndaten ω
Sukzessive Vergröberung (Entfernen schwacher Äste)	Modellstrafterm
Auswahl des besten Teilbaums	Validierungsdaten $\tilde{\omega}$

Lokaler Resubstitutionsfehler

Relative Anzahl der Fehler bei Entscheidung in β:

$$\mathrm{R}(\beta) \stackrel{\text{def}}{=} \frac{\#\text{ falsch klassifiziert in } \beta}{\#\text{ alle Objekte}} = \frac{|\omega(\beta)| - \max_\kappa |\omega_\kappa(\beta)|}{|\omega(\beta_\triangle)|}$$

Kumulativer Resubstitutionsfehler

$\mathcal{B}_\ell^\beta$ = Menge aller Blattknoten in dem von β dominierten Teilbaum

$$\mathrm{R}^*(\beta) \stackrel{\text{def}}{=} \sum_{\beta' \in \mathcal{B}_\ell^\beta} \mathrm{R}(\beta')$$

Bemerkung Es gilt für alle $\beta \in \mathcal{B}$: $\mathrm{R}^*(\beta) \leq \mathrm{R}(\beta)$

Strafterme versus Kreuzvalidierung

Effizienz eines Teilbaums

Gut entmischende Teilbäume werden belohnt,
aber zersplitterungsverdächtige Teilbäume werden bestraft!

$$\Delta_{\text{eff}}(\beta) \stackrel{\text{def}}{=} \frac{\text{Fehlerzuwachs in } \beta}{\#\text{ eingesparte Knoten}} = \frac{\mathrm{R}(\beta) - \mathrm{R}^*(\beta)}{|\mathcal{B}_\ell^\beta| - 1}$$

Kreuzvalidierungsfehler

Jedem Objekt $\boldsymbol{x} \in \tilde{\omega}$ wird durch einen Entscheidungsbaum ein Blattknoten $\beta(\boldsymbol{x})$ und damit auch eine Klassenmarkierung $\delta_\ell(\beta(\boldsymbol{x}))$ zugeordnet.

$$\tilde{\varepsilon}(\mathcal{B}) \stackrel{\text{def}}{=} \frac{\sum_{\kappa=1}^{K} |\{\boldsymbol{x} \in \tilde{\omega}_\kappa \mid \delta_\ell(\beta(\boldsymbol{x})) \neq \kappa\}|}{|\tilde{\omega}(\beta_\triangle)|}$$

CART Pruning-Algorithmus

Breiman, Friedman, Olshen & Stone (1984)

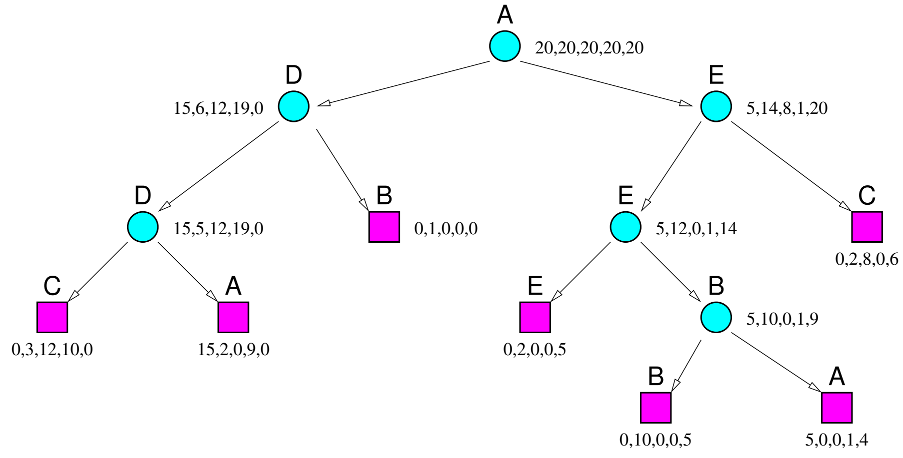
(Algorithmus)

1. ZÜCHTEN
   Expandiere initialen Baum $\mathcal{B}^{(0)}$ mittels Lerndaten $\omega_1, \ldots, \omega_K$ unter Einhaltung des *„Reinheitsgebotes"*.
2. SUKZESSIVES ZURÜCKSCHNEIDEN
   Erzeuge eine Folge gestutzter Teilbäume von $\mathcal{B}^{(0)}$
   - a. Setze $i \to 0$.
   - b. Berechne alle Effizienzwerte $\Delta_{\text{eff}}(\beta)$, $\beta \in \mathcal{B}^{(i)}$.
   - c. Wähle Knoten $\beta^* \in \mathcal{B}^{(i)}$ mit minimaler Effizienz.
   - d. Kappe den Teilbaum unterhalb β^*.
   - e. Setze $i \leftarrow i + 1$ und bezeichne gekürzten Baum als $\mathcal{B}^{(i)}$.
   - f. Ist $\mathcal{B}^{(i)} \neq \{\beta_\triangle\}$, dann $\rightsquigarrow$ b.
3. AUSWAHL NACH VALIDIERUNGSFEHLER
   Wähle aus $\{\mathcal{B}^{(i)} \mid i = 0, 1, 2, \ldots\}$ denjenigen Baum mit geringstem Fehler auf den Validierungsdaten $\tilde{\omega}_1, \ldots, \tilde{\omega}_K$.

(Algorithmus)

Beispiel — CART-Algorithmus

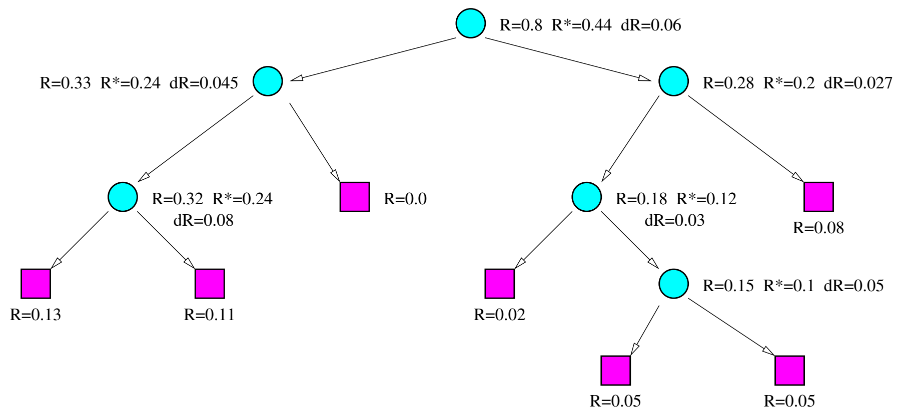
5 Klassen · 6 Fragen · 7 Blätter · 100 Objekte



Klassenhäufigkeiten je Knoten
Bestklassenmerkierung je Knoten

Beispiel — CART-Algorithmus

5 Klassen · 6 Fragen · 7 Blätter · 100 Objekte



Lokale Resubstitutionsfehlerraten
Kumulative Resubstitutionsfehlerraten
Effizienzen — *nur innere Knoten werden gezählt*

Kreuzvalidierendes Stutzen der Äste

„Frühe Validierung" — schon zur Bewertung statt erst zur Auswahl

Lokale Fehlerrate

im Knoten β nach Einschleusen der Konterdaten ω:

$$\varepsilon(\beta) \stackrel{\text{def}}{=} 1 - \frac{|\omega_\kappa(\beta)|}{|\omega(\beta)|} \quad \text{mit } \kappa := \delta_\ell(\beta) \text{ oder } \kappa := \underset{\lambda}{\operatorname{argmax}} |\omega_\lambda(\beta)|$$

Kumulative Fehlerrate

nach Durchschleusen von ω bis zu den Blättern:

$$\varepsilon^\star(\beta) \stackrel{\text{def}}{=} \sum_{\beta' \in \mathcal{B}_\ell^\beta} \frac{|\omega(\beta')|}{|\omega(\beta)|} \cdot \varepsilon(\beta')$$

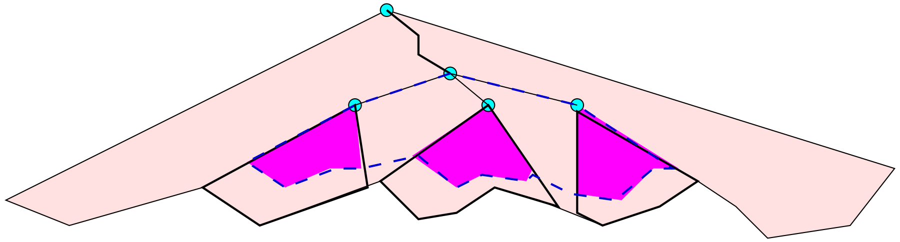
Die **Gesamtfehlerrate** ist $\varepsilon(\mathcal{B}) = \varepsilon^\star(\beta_\triangle)$

Minimale Fehlerrate

aller Teilbäume $\mathcal{B}^\beta$ unterm Knoten β:

$$\varepsilon^\forall(\beta) \stackrel{\text{def}}{=} \min\left\{\varepsilon(\mathcal{B}') \mid \mathcal{B}' \text{ Teilbaum von } \mathcal{B}^\beta\right\}$$

Induktive Bottom-Up Beschneidung

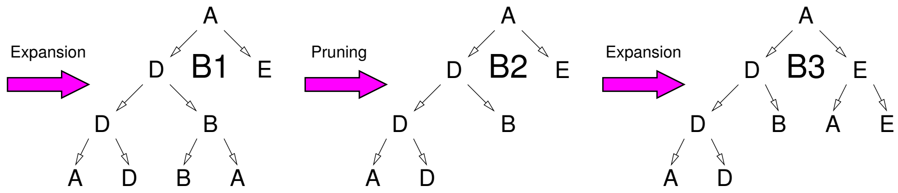


Lemma

Sei $(\mathcal{B}, Q, \delta_\ell)$ ein binärer Entscheidungsbaum über $\boldsymbol{\Omega} = \mathbb{R}^D$. Die optimale Fehlerrate des Teilbaums $\mathcal{B}^\beta$ berechnet sich nach folgender Rekursion:

$$\varepsilon^\forall(\beta) = \begin{cases} \varepsilon(\beta) & \beta \in \mathcal{B}_\ell \\ \min \begin{pmatrix} \varepsilon(\beta) \\ \sum\limits_{\beta \prec \beta'} p(\beta'|\beta) \cdot \varepsilon^\forall(\beta') \end{pmatrix} & \beta \notin \mathcal{B}_\ell \end{cases}$$

Wiederholtes Züchten und Beschneiden



Expansionsphase · *top-down*

1. Schleuse die Daten ω^a bis zu den Blattknoten von $\mathcal{B}^{(i)}$.
2. Bestimme die Mengen $\omega^a_\kappa(\beta)$ für alle κ, β.
3. Züchte für alle Blattknoten $\beta \in \mathcal{B}^{(i)}_\ell$ einen Teilbaum unter β mittels $\omega^a(\beta)$.

Pruningphase · *bottom-up*

1. Schleuse die Daten ω^a bis zu den Blattknoten von $\mathcal{B}^{(i)}$.
2. Markiere alle $\beta \in \mathcal{B}^{(i)}$ mit neuen Klassen $\delta_\ell(\beta)$
3. Überprüfe alle β durch Vergleich von lokaler und minimaler RFR auf *Eliminierbarkeit*.

Gelfands IEP-Algorithmus

„Iterative Expansion-Pruning"

(Algorithmus)

1. INITIALISIERUNG
   Setze $i \leftarrow 0$ und $\mathcal{B}^{(0)} \leftarrow \{\beta_\triangle\}$.
2. ERSTES EXPANDIEREN
   Expandiere $\mathcal{B}^{(i)}$ mit den Daten ω^a. $\leadsto \mathcal{B}^{(i+1)}$
3. ERSTES STUTZEN
   Beschneide $\mathcal{B}^{(i+1)}$ mit den Daten ω^b. $\leadsto \mathcal{B}^{(i+2)}$
4. ZWEITES EXPANDIEREN
   Expandiere $\mathcal{B}^{(i+2)}$ mit den Daten ω^b. $\leadsto \mathcal{B}^{(i+3)}$
5. ZWEITES STUTZEN
   Beschneide $\mathcal{B}^{(i+3)}$ mit den Daten ω^a. $\leadsto \mathcal{B}^{(i+4)}$
6. TERMINIERUNG
   Falls $\mathcal{B}^{(i+2)} \equiv \mathcal{B}^{(i+4)}$, dann $\leadsto$ ENDE.
7. WIEDERHOLUNG
   Setze $i \leftarrow i + 4$ und weiter bei $\leadsto$ 2.

(Algorithmus)

Disjunkt: $\left\{ \begin{array}{l} \omega^a = \omega^a_1 \cup \omega^a_2 \cup \ldots \cup \omega^a_K \\ \omega^b = \omega^b_1 \cup \omega^b_2 \cup \ldots \cup \omega^b_K \end{array} \right.$

Die Auswahl der besten Frage

Monothetische Knoten $\leadsto$ keine Attributkombinationen

Problem

*Die Expansion eines jeden Knotens β im TDI-Algorithmus erfordert die $\Delta_Q(\beta)$-Bewertung **jeder Frage** Q zu **jedem Attribut** $\mathcal{X}_n$!*

Nominale Attribute

Wieviele Zwei- oder Mehrwege-Fragen sind zu testen?

- Wertverzweigung — **eine** Frage/Attribut
- Attribut-Wert-Gleichung — $|\mathcal{X}_n|$ Targets/Attribut
- **Literalkomplex** — $2^{|\mathcal{X}_n|}/2$ Mengen/Attribut

Numerische und ordinale Attribute

Wieviele Schwellenwert-Fragen sind zu testen?

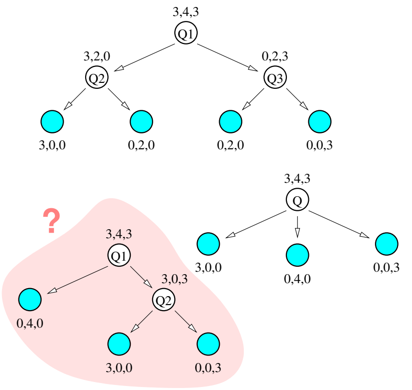
- Ordinale Attribute — $|\mathcal{X}_n| - 1$ Schwellen/Attribut
- **Numerische Attribute** — $|\omega(\beta)| - 1$ Schwellen/Attribut

Die Befragung nominaler Attribute

Symmetrische Verzweigung $x_n = ?$ versus asymmetrische Verzweigung $x_n \stackrel{?}{=} \xi_\ell$

Datenfragmentierung

Die minimal zersplitternde Folge **binärer** Fragen wird nicht automatisch gefunden.



Unbalancierte Auswahl

Die Maximierung der Entscheidungssicherheit bevorzugt systematisch Fragen mit **hohem Verzweigungsfaktor**.

Gain Ratio Impurity

Abhilfe schafft Normierung auf die maximale Entropie:

$$\Delta'_Q(\beta) \stackrel{\text{def}}{=} \frac{\Im(\beta) \; - \; \sum_{j=1}^{L} p_j \cdot \Im(\beta_j)}{\mathcal{H}(p_1, \ldots, p_L)}$$

Literalkomplexe in Zweiklassen-Szenarien

Auswahl der besten Teilmenge

Aufgabenstellung

Finde zum Attribut $\mathcal{X}_n$ in β diejenige Teilmengenfrage

$$Q : x_n \;\mapsto\; \begin{cases} 1 & x_n \in U \\ 0 & x_n \notin U \end{cases}, \qquad U \subset \mathcal{X}_n$$

mit der max. Reduktion $\Delta_Q(\beta)$ der Entscheidungsunsicherheit.

Premiumschlitten & Volumenmodelle

Objekte = Fahrzeuge · Klassen Ω_1 und Ω_2 · Attribut x_{19} (Hersteller)

$\mathcal{X}_{19}$	VW	Benz	Alfa	Dacia	BMW	Porsche
Ω_1	112	9	3	1	28	5
Ω_2	112	1	2	4	12	0
$\hat{\mathrm{P}}(1\|\xi)$	0.5	0.9	0.6	0.2	0.7	1.0

Aufsteigende $\hat{\mathrm{P}}(1|\xi)$-Sortierung ➪ nur 5 mögliche **optimale** Fragen:

$x_{19} \in \{\text{Dacia}\}$ $\quad x_{19} \in \{\text{Dacia, VW}\}$ $\quad x_{19} \in \{\text{Dacia, VW, Alfa}\}$ $\quad x_{19} \in \{\text{Dacia, VW, Alfa, BMW}\}$ $\quad x_{19} \in \{\text{Dacia, VW, Alfa, BMW, Benz}\}$

Der Zwillingssatz („*Twoing Theorem*“)

Linearer Suchaufwand für entropiegesteuertes Zweiklassen-Lernen

Satz

Es sei $\mathcal{X}_n = \{\xi_1, \ldots, \xi_L\}$ der (nominale) Wertebereich des n-ten Attributs, und es zerfalle die Lernstichprobe $\omega \subset \mathbf{\Omega}$ in zwei Klassenbereiche ω_1, ω_2. Mit den Bezeichnungen

$$\hat{\mathrm{P}}(\kappa|\xi_\ell) \stackrel{def}{=} \frac{|\{\mathbf{x} \in \omega_\kappa \mid x_n = \xi_\ell\}|}{|\{\mathbf{x} \in \omega \mid x_n = \xi_\ell\}|}$$

für $\kappa = 1, 2$ und $\ell = 1, \ldots, L$ seien infolge geeigneter Sortierung der ξ_ℓ die Häufigkeitsbeziehungen

$$\hat{\mathrm{P}}(1|\xi_1) \;\le\; \hat{\mathrm{P}}(1|\xi_2) \;\le\; \ldots \;\le\; \hat{\mathrm{P}}(1|\xi_L)$$

gültig. Dann besitzt die Teilmengenfrage mit der maximalen Homogenitätsreduktion in Bezug auf das Entropiemaß die Gestalt

$$x_n \;\in\; \{\xi_1, \ldots, \xi_\ell\}$$

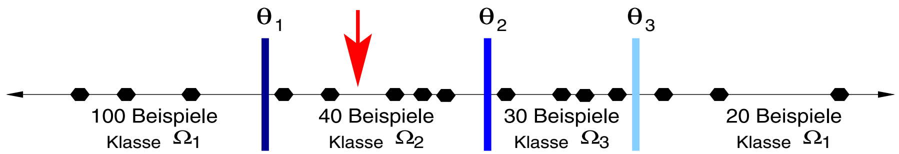
für ein geeignetes ℓ mit $1 < \ell < L$.

Schwellenwertfragen

Numerische und ordinale Attribute · zwei oder mehr Klassen

Reduzierter Suchaufwand für $\theta \in \mathcal{X}_n = \mathbb{R}$

- Nur $T_\beta = |\omega(\beta)|$ **Mittelpunktschwellen** zu prüfen.
- Nur **klassentrennende** Schwellen können $\Delta_Q(\beta)$-maximal sein.
- Es gibt eine **Rekursionsformel** für $\Delta_{Q,n,\theta}(\beta)$.



Sortierung $\mathrm{O}(T \log T)$

Aufsteigendes Sortieren der $\mathcal{X}_n$-Attributwerte in $\omega(\beta)$:

$$a_1 < a_2 < a_3 \ldots < a_t < \ldots < a_{T_\beta}$$

Mittelpunktschwellen

Suffizienter Satz von Schwellenwerten für $Q_{n,\theta}$:

$$\theta_t \;=\; \frac{a_{t+1} - a_t}{2}, \quad t = 1, 2, \ldots, T_\beta - 1$$

Separierende Schwellenwerte

Definition

Eine Mittelpunktschwelle θ_t von $\{x_n \mid \boldsymbol{x} \in \omega(\beta)\}$ heißt **innere Schwelle** von $\mathcal{X}_n$ in β, falls alle Objekte $\boldsymbol{x} \in \omega(\beta)$ mit $x_n = a_t$ oder $x_n = a_{t+1}$ zu einundderselben Klasse Ω_κ gehören.

Andernfalls heißt θ_n **separierende Schwelle** oder **Klassengrenze**.

Lemma (Fayyad & Irani, 1992)

Sind $[\theta_t]$ die Mittelpunktschwellen zur assoziierten Stichprobe $[\omega_\kappa(\beta)]$ von β zum Attribut $\mathcal{X}_n$, und gilt

$$\theta_{t^\star} \;=\; \underset{\theta_t}{\mathit{argmax}}\, \Delta_{\{x_n \leq \theta_t\}}(\beta)$$

für die entropiebezogene Entscheidungsunsicherheit, so ist $\theta_{t^\star}$ notwendigerweise eine Klassengrenze.

Bemerkung

Je stärker sich die Objekte klassenweise auf der $\mathcal{X}_n$-Achse häufen, desto weniger Reduktionswerte müssen berechnet werden.

Inkrementelle $\Delta_Q(\beta)$-Berechnung

Lemma

Es seien $\theta_1 < \theta_2$ zwei benachbarte Klassengrenzen für $\mathcal{X}_n$ in $\omega(\beta)$, zwischen denen genau m Muster der Klasse Ω_κ liegen. Dann gilt die Rekursionsformel

$$\Delta_{\{x_n \leq \theta_2\}}(\beta) \;=\; \Delta_{\{x_n \leq \theta_1\}}(\beta) \;+\; \frac{h(\ell, r) - h(\ell + m, r + m) + h(\ell_\kappa + m, r_\kappa + m) - h(\ell_\kappa, r_\kappa)}{T}$$

mit den Abkürzungen

$$h(p, q) \;=\; p \log_2 p - q \log_2 q$$

und den Zählwerten

$$\begin{aligned} \ell_\kappa &= |\{x_d < \theta_1 \mid \boldsymbol{x} \in \omega_\kappa\}| & \qquad \ell &= \textstyle\sum_\kappa \ell_\kappa \\ r_\kappa &= |\{x_d > \theta_1 \mid \boldsymbol{x} \in \omega_\kappa\}| & \qquad r &= \textstyle\sum_\kappa r_\kappa \end{aligned}$$

Attribute mit Fehlanzeigen

Imputation

Wenn $x_n = ?$, so setze einen Standardwert $\hat{\xi}$ ein.

- Wähle für $\hat{\xi}$ das globale Attributmittel μ_n.
- Wähle für $\hat{\xi}$ das **lokale** Attributmittel $\mu_n(\beta)$.

Überlagerung

Wenn $x_n = ?$, so folge in der Abrufphase parallel allen Verzeigungen.

- Während der Lernphase werden defiziente Objekte bei der $\Delta_Q(\beta)$-Berechnung **ignoriert** oder **pejorisiert**.

Surrogate Split

Wenn $x_i = ?$, so beantworte in der Abrufphase die/eine Ersatzfrage.

- In der Lernphase merkt man/frau sich die besten Fragen zum zweitbesten Attribut (ggf. weitere Alternativen).

Polythetische Entscheidungsfragen

Über das Züchten „*schiefer*“ statt achsenparalleler Entscheidungsbäume

Attributübergreifende Dichotomien (linear)

$$a_0 + \sum_{i=1}^{N} a_i x_i \;\overset{?}{\leq}\; 0$$

Trennfunktionsparameter mit guter Klassenentmischung!

- **CART/LC**
  Gradientenabstieg via $\Delta_{\boldsymbol{a}}(\beta)$
- **SADT**
  Simulated Annealing of Decision Trees
- **LMDT**
  Linear Machine Decision Trees („*ADALINE-Knoten*“)
- **QUEST**
  Multivariate Variante des QUEST-Algorithmus

QUEST-Algorithmus

Quick Unbiased Efficient Statistical Tree

1. KLASSENBEDINGTE MITTELWERTE
   Berechne β-lokale klassenbezogene Mittelwertvektoren $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$.
2. STATISTISCHER HYPOTHESENTEST
   Fisher-Test für die Nullhypothesen
   $$H_0 : \ \mu_1^{(n)} = \mu_2^{(n)} = \ldots = \mu_K^{(n)}$$
3. ZENTREN CLUSTERN („*2-means*“)
   Partitioniere für das Gewinnerattribut $n^* \in \{1 : N\}$ die K Mittelwerte $\mu_1^{(n^*)}, \mu_2^{(n^*)}, \ldots, \mu_K^{(n^*)}$.
4. KONSTRUIERE TRENNFRAGE
   Berechne NV-Dichteparameter für die beiden Cluster.
   $$\mathcal{N}(x_{n^*} \mid \mu_1, \sigma_1^2) \quad \overset{?}{\leq} \quad \mathcal{N}(x_{n^*} \mid \mu_2, \sigma_2^2)$$

Klassenprädiktoren

in den inneren und den Blattknoten des Entscheidungsbaumes

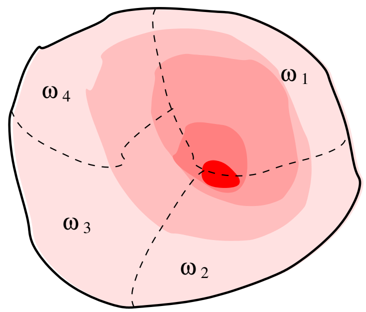
Schleusungspfad

Jedes Objekt $\boldsymbol{x} \in \mathcal{X}$ beschreibt einen Pfad

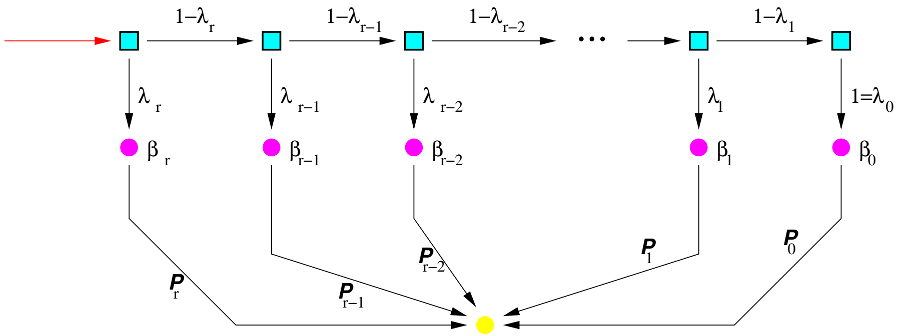
$$\beta_\triangle = \beta_0(\boldsymbol{x}) \prec \beta_1(\boldsymbol{x}) \prec \beta_2(\boldsymbol{x}) \prec \ldots \prec \beta_{r-1}(\boldsymbol{x}) \prec \beta_r(\boldsymbol{x}) = \beta(\boldsymbol{x})$$

Lokale Prädiktoren

$\hat{p}_\kappa(\beta_0)$	$=$	$\lvert\omega_\kappa\rvert$	$/$	$\lvert\omega\rvert$
$\wr\wr$		$\vee\!\mid$		$\vee\!\mid$
$\hat{p}_\kappa(\beta_1)$	$=$	$\lvert\omega_\kappa(\beta_1)\rvert$	$/$	$\lvert\omega(\beta_1)\rvert$
$\wr\wr$		$\vee\!\mid$		$\vee\!\mid$
$\hat{p}_\kappa(\beta_2)$	$=$	$\lvert\omega_\kappa(\beta_2)\rvert$	$/$	$\lvert\omega(\beta_2)\rvert$
$\wr\wr$		$\vee\!\mid$		$\vee\!\mid$
$\cdots$		$\cdots$		$\cdots$
$\wr\wr$		$\vee\!\mid$		$\vee\!\mid$
$\hat{p}_\kappa(\beta_r)$	$=$	$\lvert\omega_\kappa(\beta_r)\rvert$	$/$	$\lvert\omega(\beta_r)\rvert$



Lineare Interpolation von Klassenprädiktoren



Interpolationsformel für a posteriori-Klassenwahr’keiten

Maximum-Likelihood-Koeffizienten nach EM-Algorithmus

$$\tilde{p}_\kappa(\beta_r) \ = \ \begin{cases} 1 \cdot \hat{p}_\kappa(\beta_\triangle) & r = 0 \\ \lambda_r \cdot \hat{p}_\kappa(\beta_r) \ + \ (1-\lambda_r) \cdot \tilde{p}_\kappa(\beta_{r-1}) & r > 1 \end{cases}$$

Leo Breimans Random Forests

Zweifache Ensembletechnik: Objekte (*bagging*) & Attribute

Lernprobe ω, Wälder $M \in \mathbb{N}$, Auswahl $T_b \leq |\omega|$ und $N_b \ll N$.

1. LERNPHASE
   Erzeuge Bäume $\mathcal{B}^{(m)}$, $m = 1, \ldots, M$:
   - Lernprobe $\omega^{(m)} \subset \omega$ via T_b-**Bootstrap** **mit** Ersetzen
   - Zufallsbaum $\mathcal{B}^{(m)}$ via TDI-Algorithmus
   - EINGESCHRÄNKTE LOKALE FRAGEAUSWAHL:
     $\mathcal{A}_\beta \subset \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$, $|\mathcal{A}_\beta| = N_b$ via N_b-**Bootstrap** **ohne** Ers.
   - Kein Zurückstutzen!
2. ABRUFPHASE
   Mehrheitsentscheidung unter allen Bäumen des Waldes:
   $$\kappa^*(\boldsymbol{x}) \ = \ \text{argmax}_\kappa \left|\left\{\mathcal{B}^{(m)} \mid \delta_\ell(\beta^{(m)}(\boldsymbol{x})) = \kappa\right\}\right|$$

Bemerkung

Pro: Effizient, skalierbar (N, T), exzellentes Erkennungsverhalten.
Contra: Überanpassung, Reproduzierbarkeit, Präferenz stufenreicher Nominalattribute.

Vorhersage Konzeptlernen Versionenräume Bayesregel Regression Logitmodell Präferenzen CART Σ

Zusammenfassung (4)

1. Der **Konzeptraum** enthält die **zu lernenden**, der **Hypothesenraum** die **lernbaren** Teilmengen des Objektraums.
2. Der **Versionenraum** besteht aus allen **konsistenten** Hypothesen und ist als **Halbordnungsintervall** darstellbar.
3. Die Hypothesen des **Sterns** grenzen ein Positivbeispiel gegen alle Negativbeispiele ab.
4. **Lineare Diskriminanten** approximieren die **ideale Trennfunktion** im Quadratmittelsinn.
5. **Loglineare Diskriminanten** approximieren die **a posteriori** Klassenwahr'keiten.
6. Beide Lernverfahren lassen sich **regularisieren** und **dualisieren**.
7. **Entscheidungsbäume** klassifizieren durch hierarchische Befragung **numerischer & diskreter** Attribute.
8. Sie werden durch ein **gieriges Top-Down-Verfahren** aus den Daten gelernt.
9. Für die lokale Suche nach der maximal **klassenentmischenden** Frage gibt es effiziente Verfahren.