

Literatur Hastie, Tibshirani and Friedman. The Elements of Statistical Learning. 2nd ed., Springer 2009., Kap. 4 ,15

Fahrmeir, Hammerle und Tutz: Multivariate statistische Verfahren. De Gruyter, 1996. Kap. 8

Problemstellung Zuordnung von Untersuchungseinheiten zu Klassen $Y = 1, \dots, k$ aufgrund des Datenvektors \mathbf{x} .

Beispiele:

- Kreditwürdigkeit und Persönlichkeitsmerkmale
- Tierarten und Merkmale
- Krankheitsstatus und diagnostische Merkmale

Einfache Strategien für $Y \in \{0,1\}$

$$\delta(x) = 1 \Leftrightarrow \hat{Y} = 1 \text{ (Entscheidung für } Y=1\text{)}$$

1. Lineare Regression von Indikatorvariablen

Y auf X

$$Y = x'\beta + \varepsilon$$

$$\delta(x) = 1 \Leftrightarrow \hat{Y} = 1 \Leftrightarrow x'\hat{\beta} > 0.5$$

\Rightarrow Grenze ist linear in x .

2. K -nächste Nachbarn (K -Nearest Neighbor Classifier)

Betrachte K Beobachtungen als Nachbarn $N_K(x)$ von x

$$\frac{1}{K} \sum_{x_i \in N_K(x)} y_i > \frac{1}{2} \Rightarrow \delta(x) = 1$$

\Rightarrow Grenze stark abhängig von den Daten und der Wahl von K .

3. Logistische Regression

Modell

$$P(Y = 1|x) = G(x'\beta) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

$$G(t) = (1 + \exp(-t))^{-1} = \frac{\exp(t)}{1 + \exp(t)}$$

$$\delta(x) = 1 \Leftrightarrow G(x'\hat{\beta}) > \frac{1}{2} \Leftrightarrow x'\hat{\beta} > G^{-1}\left(\frac{1}{2}\right)$$

- keine Annahme über die Verteilung von $x|y$
- keine Annahme zur a-priori W'keit

- Prognoseproblem
- Supervised Learning
- Lerndatensatz mit gegebener Zuordnung
→ Ableitung von Regeln zur Zuordnung neuer Einheiten

Bias Variance trade off

Unterscheidung in

- Lerndaten
- Testdaten

Höher Modellkomplexität führt zu

- Besserer Anpassung im Lerndatensatz (Bias), höherer Variabilität
- zunächst bessere Prognosegüte, aber dann wg. Overfitting zu schlechter Prognosegüte

Allgemeiner Ansatz

Verlustfunktion

$$L(y, \delta(x))$$

Ziel: Minimiere Erwarteten Verlust

Aus

$$L(y, \delta(x)) = 1 - I(\delta(x) = y)$$

folgt: $E(L)$: Fehlklassifikationswahrscheinlichkeit

Bayes-Zuordnung

Wichtige Größen für das Klassifikationsproblem sind

- die *a priori*-Wahrscheinlichkeiten $p(r) = P(Y = r), r = 1, \dots, k,$
- die *a posteriori*-Wahrscheinlichkeiten
 $P(r | x) = P(Y = r | x), r = 1, \dots, k,$
- die *Verteilung der Merkmale*, gegeben die Klasse, bestimmt durch die Dichten $f(x | 1), \dots, f(x | k).$
- die *Mischverteilung* der Population
 $f(x) = f(x | 1)p(1) + \dots + f(x | k)p(k).$

Satz von Bayes:

$$P(r | x) = \frac{f(x | r)p(r)}{f(x)} = \frac{f(x | r)p(r)}{\sum_{i=1}^k f(x | i)p(i)}$$

Bayes-Zuordnung

Regel:

$$\delta^*(x) = r \iff P(r | x) = \max_{i=1,\dots,k} P(i | x)$$

Dabei sei $P(i|x)$ gegeben.

- *Wahrscheinlichkeit einer Fehlklassifikation, gegeben der feste Merkmalsvektor x*

$$\begin{aligned}\varepsilon(\mathbf{x}) &= P(\delta(x) \neq Y | x) = 1 - P(\delta(x) = Y | x) \\ &= 1 - P(\delta(x) | x).\end{aligned}$$

- *Verwechslungswahrscheinlichkeit oder individuelle Fehlerrate*

$$\varepsilon_{rs} = P(\delta(x) = s | Y = r) = \int_{x:\delta(x)=s} f(x | r) dx$$

- Globale Fehlklassifikationswahrscheinlichkeit oder Gesamt-Fehlerrate

$$\varepsilon = P(\delta(x) \neq Y)$$

- Wahrscheinlichkeit einer Fehlklassifikation, gegeben das Objekt entstammt der Klasse r

$$\varepsilon_r = P(\delta(x) \neq r \mid Y = r) = \sum_{s \neq r} \varepsilon_{rs}$$

Es gilt:

$$\begin{aligned} \varepsilon &= P(\delta(x) \neq Y) = \sum_{r=1}^k P(\delta(x) \neq r \mid Y = r) p(r) = \sum_{r=1}^k \varepsilon_r p(r) = \sum_{r=1}^k \sum_{s \neq r} \varepsilon_{rs} p(r) \\ &= P(\delta(x) \neq Y) = \int P(\delta(x) \neq Y \mid x) f(x) dx = \int \varepsilon(x) f(x) dx \end{aligned}$$

Optimalität der Bayes–Zuordnung

Die Bayes–Zuordnung

$$\delta^*(x) = r \iff P(r \mid x) = \max_{i=1,\dots,k} P(i \mid x)$$

minimiert die Gesamtfehlerrate ε .

Äquivalente Diskriminanzfunktionen:

- a) $d_r(x) = P(r \mid x),$
- b) $d_r(x) = f(x \mid r)p(r)/f(x).$
- c) $d_r(x) = f(x \mid r)p(r).$
- d) $d_r(x) = \log(f(x \mid r)) + \log(p(r)).$

Kostenoptimale Bayes–Zuordnung:

$c(i, j) = c_{ij}$ = Kosten der Zuordnung eines Objekts aus Klasse i in die Klasse j .

Es gelte $c_{ii} = 0$.

Bedingtes Risiko, gegeben x

$$r(x) = \sum_{i=1}^k c_{i, \delta(x)} P(i \mid x).$$

Individuelles Risiko

$$r_{ij} = c_{ij} P(\delta(x) = j \mid Y = i) = c_{ij} \int_{x: \delta(x)=j} f(x \mid i) dx$$

Risiko, gegeben Klasse i

$$r_i = \sum_{j=1}^k r_{ij}.$$

Kostenoptimale Bayes–Zuordnung:

Bayes–Risiko

$$R = E(c_{Y, \delta(x)}) = \sum_{i=1}^k r_i p(i) = \int r(x) f(x) dx.$$

Bayes–Zuordnung mit Kosten

$$\delta^*(x) = r \iff \sum_{i=1}^k P(i | x) c_{ir} = \min_{j=1, \dots, k} \sum_{i=1}^k P(i | x) c_{ij}$$

Mit den Diskriminanzfunktionen

$$d_r(x) = - \sum_{i=1}^k P(i | x) c_{ir},$$

erhält man

$$\delta^*(x) = r \iff d_r(x) = \max_{i=1, \dots, k} d_i(x).$$

Maximum–Likelihood–(ML)–Zuordnungsregel

$$\delta_{ML}(x) = r \iff f(x \mid r) = \max_{i=1,\dots,k} f(x \mid i)$$

ML-Regel entspricht Bayes-Regel für gleiche a-priori-W'keiten.

Zuordnungsregel unter Normalverteilung

(a) Klassenverteilungen ($i \in \{1, \dots, k\}$):

$$f(x) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_i)^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right\}$$

Diskriminanzfunktion:

$$d_i(x) = \ln(f(x | i)) + \ln(p(i))$$

$$d_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) - \frac{1}{2} \ln(\det \Sigma_i) + \ln p(i), \quad i = 1, \dots, k.$$

Zuordnungsregel unter Normalverteilung

Bayes–Regel:

$$\delta(x) = r \iff d_r(x) = \max_{i=1,\dots,k} d_i(x)$$

(b) *Zusätzlich unabhängige standardisierte Merkmale, $\Sigma_i = \sigma^2 \mathbf{I}$:*

$$d_i(x) = -\frac{\|x - \mu_i\|^2}{2\sigma^2} + \ln p(i)$$

Sind die a priori Wahrscheinlichkeiten gleich groß, so reduziert sich die Diskriminanzfunktion auf

$$d_i(x) = -\|x - \mu_i\|^2.$$

Zuordnungsregel unter Normalverteilung

(c) *Klassenweise identische Kovarianzmatrizen; $\Sigma_i = \Sigma$, $i = 1, \dots, k$:*

$$d_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + \ln p(i)$$

$$d_i(x) = \mu_i^T \Sigma^{-1} x - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln p(i), \quad i = 1, \dots, k$$

Für $k = 2$ ($Y \in \{1, 2\}$) erhält man:

$$d(x) = d_1(x) - d_2(x) = \left[x - \frac{1}{2}(\mu_1 + \mu_2) \right]^T \Sigma^{-1}(\mu_1 - \mu_2) - \ln \frac{p(2)}{p(1)}$$

$$\delta(x) = 1 \iff d_1(x) > d_2(x) \iff d(x) > 0$$

Zuordnungsregel unter Normalverteilung

Schätzer für μ_i und Σ :

$$\hat{\mu}_i = \bar{x}_i, \quad \hat{\Sigma} = S = \frac{1}{N - k} \sum_{i=1}^k \sum_{n=1}^{N_i} (x_{in} - \bar{x}_i)(x_{in} - \bar{x}_i)^T$$

Fehlerraten im Zwei-Gruppen-Fall: ($\Sigma_i = \Sigma$)

Bei Verwendung der Bayes-Regel ergibt sich die Gesamtfehlerrate

$$\varepsilon = p(1)\varepsilon_{12} + p(2)\varepsilon_{21}$$

mit den individuellen Fehlerraten

$$\varepsilon_{12} = \Phi\left(\frac{\ln[p(2)/p(1)] - \delta^2/2}{\delta}\right), \quad \varepsilon_{21} = \Phi\left(\frac{-\ln[p(2)/p(1)] - \delta^2/2}{\delta}\right).$$

Fehlerraten im Zwei-Gruppen-Fall: ($\Sigma_i = \Sigma$)

Mahalanobis-Distanz

$$\delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2).$$

$$\hat{d}(x) = [x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]^T S^{-1}(\bar{x}_1 - \bar{x}_2) - c$$

$c = \ln\{p(2)/p(1)\}$ bzw. $c = 0$, ergibt sich

$$\varepsilon_{12} = \Phi\left(-\frac{\hat{d}(\mu_1)}{\sqrt{\text{var } \hat{d}(x)}}\right), \quad \varepsilon_{21} = \Phi\left(-\frac{\hat{d}(\mu_2)}{\sqrt{\text{var } \hat{d}(x)}}\right).$$

Zur Berechnung benötigt man μ_1, μ_2 und Σ !

Fishersche Diskriminanzanalyse

Daten: $x_{i1} \dots x_{in_i}$ in der Klasse i

Grundprinzip (Zwei-Klassen-Fall, $Y \in \{1, 2\}$):

Projektion $y = a^T x$ mit $\|a\| = 1$, so dass das folgende Kriterium maximiert wird

$$Q(a) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_1^2 + s_2^2},$$

wobei

$$\bar{y}_i = \sum_{j=1}^{n_i} a^T x_{ij} = a^T \bar{x}_i \quad \text{Klassenmittelpunkt,}$$

$$s_i^2 = \sum_{j=1}^{n_i} (a^T x_{ij} - a^T \bar{x}_i)^2 \quad \text{Quadratische Abweichungen in der } i\text{-ten Klasse.}$$

Fishersche Diskriminanzanalyse

Es gilt: $s_1^2 + s_2^2 = a' W a$

mit

$$W = (n_1 + n_2 - 2) S$$

$$S = \sum_{i=1}^2 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T.$$

W: Inner-Klassen-SSR-Matrix

S: gepoolte empirische Varianz-Kovarianz-Matrix

Fishersche Diskriminanzanalyse

Lösung

$$Q(a) = \frac{(a^T(\bar{x}_1 - \bar{x}_2))^2}{a^T W a} \rightarrow \max_{a \neq 0}$$

$$\frac{\partial Q(a)}{\partial a} = \frac{2a^T \cdot (\bar{x}_1 - \bar{x}_2)^2 a^T W a - 2W a (a^T(\bar{x}_1 - \bar{x}_2))^2}{(a^T W a)^2} = 0$$

$$\Rightarrow \bar{x}_1 - \bar{x}_2 = W a \frac{(a^T(\bar{x}_1 - \bar{x}_2))}{(a^T W a)}$$

$$\Rightarrow a = W^{-1}(\bar{x}_1 - \bar{x}_2) * c, \quad c = \text{Konstante}$$

Diskriminanzregel

Neue Beobachtung x

$$y := a^T x$$

$$\delta(x) = 1 \Leftrightarrow |y - \bar{y}_1| < |y - \bar{y}_2|$$

$$\delta(x) = 1 \Leftrightarrow y > \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

Es gilt $\bar{y}_1 > \bar{y}_2$

$$\delta(x) = 1 \Leftrightarrow \left(y - \frac{1}{2}(\bar{y}_1 + \bar{y}_2)\right) > 0$$

$$\Leftrightarrow (\bar{x}_1 - \bar{x}_2)^T W^{-1} \left[x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2) \right] > 0$$

Entspricht ML-Regel ($p(2)=p(1)$) mit identischen Kovarianz-Matrizen.

Mehr-Klassen-Fall ($i \in \{1, \dots, g\}$):

Kriterium: Suche $y = a^T x$ mit:

$$Q(a) = \frac{\sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^g s_i^2} \rightarrow \max$$

Lösung ergibt Eigenwertproblem:

$$W^{-1} B a = \lambda a$$

mit

$$B = \sum_{i=1}^g n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

$$W = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T.$$

Mehr-Klassen-Fall ($i \in \{1, \dots, g\}$):

Lösung des Eigenwertproblems liefert:

$$\lambda_k = \frac{a_k^T B a_k}{a_k^T W a_k} \quad a_k : \text{Eigenvektoren } k = 1, \dots, g$$

$y_k = a_k^T x$ "kanonische Variablen"

Klassifikation:

Wähle Klasse i so, dass

$$\sum_{k=1}^q a_k^T (x - \bar{x}_i)^2 \quad \text{minimal über } i$$

Klassifikationsbäume und Regressionsbäume (CART)

CART ist eine der einfachen Techniken in der Klassifikation.

Der CART Algorithmus generiert einen Klassifikationsbaum (Entscheidungsbaum) über binäre Aufteilungen der Daten.

Splits werden jeweils nur über eine Variable vorgenommen.

Wahl der Splits so, dass die beiden Untermengen eines Splits bzgl. der Klassifikationsvariablen möglichst homogen sind.

Bei Regressionsbäumen wird meist die Summe der quadratischen Abweichungen der Zielvariablen vom jeweiligen Mittelwert der Splits minimiert.

Split-Kriterium bei Klassifikationsbäumen

Sei die Zielvariable Y kategoriell mit Werten $k = 1, 2, \dots, K$. Für eine Partition R_m mit N_m Beobachtungen ist

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \quad i = 1, \dots, N$$

der Anteil an Beobachtungen von Klasse bzw. Kategorie k im Knoten m . Die Zuordnung des Knotens zur Klasse geschieht über $k(m) = \arg \max_k \hat{p}_{mk}$.

Impurity Maße (Maße der Unreinheit)

Übliche Maße der Unreinheit $Q_m(T)$ der Knoten m eines Baums T :

- Fehlklassifikation:

$$\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$$

- Gini Index:

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^K (1 - \hat{p}_{mk}) \hat{p}_{mk}$$

- Entropy:

$$-\sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

Impurity Maße im Fall von nur 2 Klassen

Im Fall von nur 2 Klassen (p sei der Anteil der 2. Klasse) vereinfachen sich die Maße zu:

- Fehlklassifikation: $1 - \max(p, 1 - p)$
- Gini Index: $2p(1 - p)$
- Entropy: $-p \log p - (1 - p) \log(1 - p)$

Probleme

- Stoppregel
- Instabilität
- Schätzung von Fehlern

Strategien für Stoppregel

- Stoppregel für Verbesserung des ÜnreinheitsMaßes
- Bilden eines großen Baumes und "Beschneidung" des Baumes nach einem bestimmten Kriterium ("Pruning")

$$Q_m(T) + \alpha |T| \rightarrow \text{Minimal}$$

↑

↑

Impurity Komplexitätsmaß

Bestimmung von α durch Kreuzvalidierung

Alternative : Rekursive Partioning

- Alternative zu CART
- Carolin Strobl, James Malley and Gerhard Tutz (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random forests. *Psychological Methods* 14 (4), 323–348.
- Verwende p-Werte von Signifikanztests zur Bestimmung der Splits
- Statistische Grundlage
- *fairer* Vergleich der verschieden skalierten Variablen
- Paket party

Varianz Bias Problematik

$$MSE = Bias^2 + Var$$

- Bäume berücksichtigen komplexe Interaktionen
- Bei Bäumen wegen Instabilität hohe Varianz und tendenziell geringer Bias
- Bei B unabhängigen Wiederholungen :

$$\sigma^2 = \frac{1}{B} \sigma^2$$

- Bei B Wiederholungen mit paarweiser Korrelation ρ :

$$\sigma^2 = \rho \sigma^2 * \frac{(1 - \rho)}{B} \sigma^2$$

Bootstrap Aggregating

Idee: Instabilität von CART wird durch Bootstrap-Methodik behoben

Durchführung: Ziehe B Bootstrap-Stichproben

- in der Regel Ziehen mit Zurücklegen aus den Ursprungsdaten
- CART-Algorithmus wird für jeden Baum verwendet
- Klassenzuordnung per „Abstimmung“ durch die verschiedenen Bäume
- Neben Zuordnung erhält man auch Schätzung für W' keit

Random Forest

Weiterentwicklung des Bagging:

$$\sigma^2 = \rho \sigma^2 * \frac{(1 - \rho)}{B} \sigma^2$$

Verringere Korrelation zwischen Bootstrap - Wiederholungen Vorgehen in mehreren Schritten

- ① Wiederhole folgende Schritte B mal
 - Ziehe Bootstrap-Stichprobe aus Trainingsdaten (n Einheiten zufällig mit Zurücklegen) .
 - Erzeuge einen Baum nach der CART Methode mit folgender Modifikation: Bei jedem Split werden aus den p Merkmalen m zufällig ausgewählt. Nur diese werden zur Betrachtung des (Splits) herangezogen . Aus diesen wird dann der Split gewählt (nach Kriterium (z.B. Gini)
 - Der Baum wird voll ausgebaut bis zur minimalen Knotengröße
- ② Aus den B Bäumen wird dann die Klassifizierung per Mehrheitsentscheidung vorgenommen.

Etabliertes Verfahren mit in vielen Fällen guter Performance



Allgemeines
Im Zwei-Gruppen-Fall

$$P(\hat{Y} = 0 | Y = 1) \quad \text{und} \quad P(\hat{Y} = 1 | Y = 0)$$

Falsch Negativ Falsch Positiv

Entscheidung mit Diskriminanz-Funktion
 $d(x) > s \Rightarrow \hat{y} = 1$, s Splitpunkt

Definition

- ROC-Kurven charakterisieren das Verhalten der Diskriminanzregel für alle möglichen Schwellenwerte.
- X-Achse: "False Positives": 1-Spezifität
- Y-Achse: "True Positives": Anzahl positiver Beispiele, die als positiv klassifiziert werden. Sensitivität

AUC und GINI-Koeffizient

„Area under Curve“

$$AUC = \int_0^1 ROC(x) dx, \quad 0.5 \leq AUC \leq 1$$

$$GINI = 2 \cdot \left(AUC - \frac{1}{2} \right), \quad 0 \leq GINI \leq 1$$

Maße zur Güte-Beurteilung von **Klassifikations-Funktionen**

Umso näher die Werte bei 1 sind, desto besser ist die Anpassung.

- Bewertung von Diskriminanzregeln durch Resampling und Kreuzvalidierung. Literatur: B. Bischl, O. Mersmann, H. Trautmann, and C. Weihs. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, 20(2):249–275, 2012.
- Paket mlr und Veranstaltung introduction to machine learning <https://moodle.lmu.de/course/view.php?id=3001>
- Interpretation von Black Box Verfahren zentrales Thema