Supervised Learning: Statistische Modelle mit gegebener Zielgröße

Legende:
LHS-Variable: Left Hand Side-Variable (Variable links des Gleichheitszeichens, Zielgrösse)
RHS-Variables: Left Hand Side-Variablen (Variablen rechts des Gleichheitszeichens, Einflußgrösse)

- Lineare Regression: $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + \varepsilon$ mit $\varepsilon \sim N(0; \sigma^2)$

  LHS-Variable Y ist metrisch, die RHS-Variablen $X_1, X_2, \ldots, X_p$ beliebig skaliert [1].
  R-Code für 3 RHS-Variable x_1, x_2, x_3: `lm(y~x1+x2+x3)`

- Varianzanalyse (Anova [2])

  Lineare Regression mit ausschließlich kategorialen RHS - Variablen

- Binär logistische Regression [3] : LHS-Variable Y ist 1-0-dichotom mit $\pi = P(Y=1)$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p$$

bzw.mathematisch äquivalent

$$\pi = \frac{e^{\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p}}{1 + e^{\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p}}$$

R-Code für 3 RHS-Variable: `glm(y~x1+x2+x3,family=binomial(link="logit"))`

- Poissonregression [4]
- LHS-Variable Y mit den Werten 0, 1, 2, … heißt poissonverteilt mit $\mu > 0$, wenn $P(Y=y) = \frac{\mu^y}{y!} \cdot e^{-\mu}$
  - $\log(\mu) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p$

  mathematisch äquivalent

$$\mu = e^{\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p}$$

  R-Code für 3 RHS-Variable: `glm(y~x1+x2+x3,family=poisson(link="log"))`

- Baumanalyse, Regression [5]

  Recursive partitioning: sukzessive Zerlegung des Raums [6] der RHS-Variablen in k-dimensionale, achsenparallele Quader [7] $R_1, R_2, \ldots, R_J$ des Raums der RHS–Variablen derart, dass die "Residual Sum of Squares" der Zielgröße minimiert wird

---

[1] numerisch, metrisch, ordinal, kategorial, character; mit `class("…")` abzufragen
[2] ANCOVA, veralteter Begriff: RHS-Variable z.T. auch metrisch
[3] logistische und Poissonregression, häufig verwendete Vertreter der "generalized linear models"
[4] Synonym "Count-Regression", Besonderheit: E(Y) = Var(Y), Y = Anzahl Ereignisse (z.B. Anzahl Rezidive pro Woche, Anzahl Unfälle, etc.)
[5] englisch: "recursive partitioning", "tree-analysis", "tree-based methods"

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 = Min$$

Mit $\hat{y}_{R_j}$ = Mittel der Zielgröße Y in Quader Nr. j

- Baumanalyse, Klassifikation

  Recursive partitioning: sukzessive Zerlegung des Raums der RHS-Variablen in k-dimensionale, achsenparallele Quader R_1, R_2, …, R_J des Raums der RHS-Variablen derart, dass diese möglichst "rein" sind, d.h. ganz überwiegend nur jeweils Objekte einer Klasse [8] enthalten, d. h. $p_{jk} \approx 1$.

  Die Algorithmen minimieren den Gini-Index

$$G = \sum_{k} \hat{p}_{jk} \cdot \left(1 - \hat{p}_{jk} \right)$$

  bzw. maximieren die Entropie

$$D = \sum_{k=1}^{K} \hat{p}_{jk} \cdot \log \hat{p}_{jk}$$

  die beide für $p_{jk} \approx 1$, d.h. wenn alle Quader / Knoten (j) überwiegend nur Elemente einer Kategorie (k) enthalten, die ihrem Minimum/Maximum "0" möglichst nahekommen.

- Supportvectormachines, Klassifikation

  (1) Klassifikation durch logistische Regression versagt, wenn die Teilgesamtheiten [9] linear separabel sind.

  (2) Da lineare Separabilität "fast immer" möglich ist, wenn "erheblich" mehr Merkmale als Fälle ($p \gg n$) vorliegen (z.B. Klassifikationen mit Daten aus Microarrays), scheidet die logistische Regression [10] hierbei als Instrument der Klassifikation aus.

---

[6] Die Darstellung ist auf binäre Zerlegungen beschränkt, Verallgemeinerungen sind problemlos
[7] auch Knoten genannt
[8] binär: k = 1,2 für "infiziert", "nicht-infiziert", etc.
[9] Zielgröße = 0 oder 1
[10] das "Arbeitspferd" der Biostatistik und Epidemiologie

(3) Durch den "Kernel-trick" der Support vector machines (svm) können hier Klassifikationen mit beliebiger Anzahl p von Merkmalen (Einflussgrößen) optimiert werden (Abb. 1).

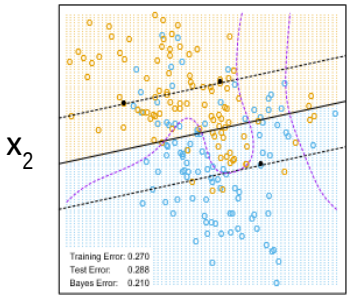

Abbildung 1

- Survivalanalyse

  LHS-Variable Y ist hier die – zensierte – zeitliche Dauer einer Episode, deren Verteilung für homogene Gesamtheiten durch Kaplan-Meier oder Nelson-Aalen geschätzt wird.

  Für inhomogene Gesamtheiten erfolgt eine Modellierung durch die relevanten Einflussgrößen, Cox-Regression oder parametrische Verfahren.

  Cox-Modell, Hazard-Funktion: $h(t, X) = h_0(t) \cdot e^{\beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p}$

  R-Code für 3 RHS-Variable: `coxph(Surv(time,cens)~x1+x2+x3)`
  `time`: zeitliche Dauer, `cens`: komplette oder abgebrochene/zensierte zeitliche Dauer

- Zeitreihenanalyse

  Abbildung 2: Trennung von Zeitreihendaten in Trend, saisonale Schwankungen und Residuum

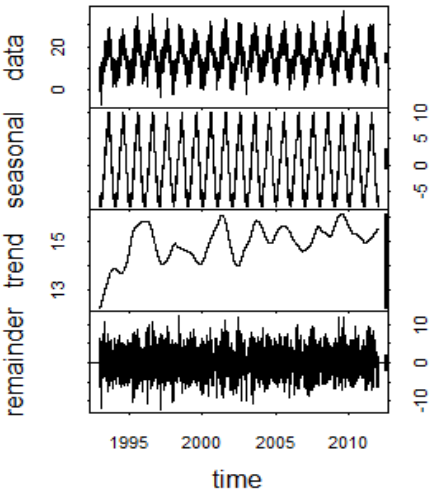

Abbildung 2