Wissensentdeckung in Datenbanken

Organisation und Überblick

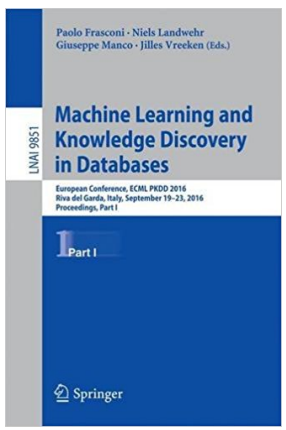
Nico Piatkowski und Uwe Ligges

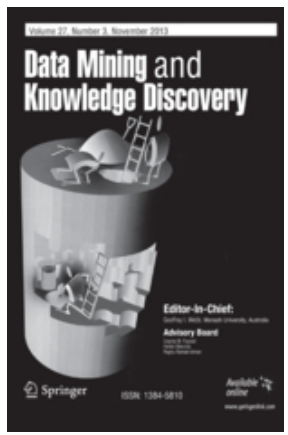
Informatik—Künstliche Intelligenz
Computergestützte Statistik
Technische Universität Dortmund

18.04.2017

Fakten

- Team
  - Vorlesung: Uwe Ligges, Nico Piatkowski
  - Übung: Sarah Schnackenberg, Sebastian Buschjäger
- 28 Termine (13/13/2)
  - Dienstags — 10:15 Uhr bis 12:00 Uhr
  - Donnerstags — 14:15 Uhr bis 16:00 Uhr
- Feiertage
  - 25.05. (Christi Himmelfahrt)
  - 15.06. (Fronleichnam)





Inhalt

18.4.	L,P	Übersicht, Einführung
20.4.	L	Statistik 1
25.4.	L	Statistik 2
27.4.	L	Stichproben, Versuchsplanung, Datenvorverarbeitung
02.5.		Optimierung, Modellklassen, Lineares Modell,
	P	Bias-Varianz, Overfitting 1
04.5.		Optimierung, Modellklassen, Lineares Modell,
	P	Bias-Varianz, Overfitting 2
09.5.	P	Data Cube, Frequent sets, Apriori, FPgrowth 1
11.5.	P	Data Cube, Frequent sets, Apriori, FPgrowth 2
16.5.	L	Einführung Klassifikation, BayesRegel, Logistische Regression
18.5.	L	kNN, Ähnlichkeitsmaße, Modellselektion
23.5.	L	Resampling, Klassifikationsbeurteilung 2

Inhalt

30.5.	P	SVM
01.6.	L	Diskriminanzanalyse (LDA) 1
06.6.	L	Diskriminanzanalyse (LDA, QDA, RDA) 2
08.6.	L	von Entscheidungsbäumen (CART) zu Wäldern
13.6.	L	Ensemble Methoden (Bagging, Boosting)
20.6.	L	Stetige Modelle Hauptkomponentenanalyse)
22.6.	P	Graphische Modelle 1
27.6.	P	Graphische Modelle 2
29.6.	P	Nicht-glatte und stochastische Optimierung
04.7.	P	Merkmalsselektion, Struktur Lernen, Regularisierung

Inhalt

06.7.	P	Clustering, k-Means, Gaussian Mixture, Latent Dirichlet Allocation 1
11.7.	P	Clustering, k-Means, Gaussian Mixture, Latent Dirichlet Allocation 2
13.7.	L	Hierarchisches Clustern; Zeitreihen 1
18.7.	L	Zeitreihen 2
20.7.	P	Künstliche Neuronale Netze 1
25.7.	P	Künstliche Neuronale Netze 2
27.7.	L,P	Zusammenfassung; Rückblick

Fragen und Kommentare

Unsere Bitte:
Mitdenken, kommentieren und Fragen stellen!

Die unterschiedliche Terminologie mag zunächst verwirren:
Merkmal, Feature, Variable, Parameter im Sprachgebrauch in Informatik, Mathematik und Statistik.

Beispiele für KDD / Data Mining

- Telekommunikationstechnik
  - riesige Datenbestände, Prozesskontrolle, Zeitreihenanalyse, Zuverlässigkeitsanalysen
- CallCenter
  - Warteschlangen, Klassifikation des Problems mit wenig Information
- Kundenkartenanalyse
  - riesige Datenbestände (und große Anzahl Variablen), Vorhersage von Kaufverhalten und Warengruppenzuordnung
- Genomik
  - riesige Datenbestände (und große Anzahl Variablen), viel Rauschen, Klassifikation zur Wissensentdeckung

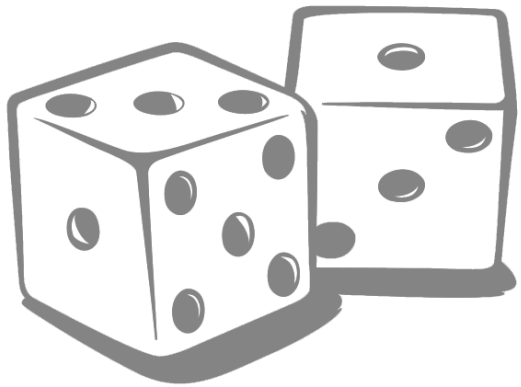
Statistische Grundlagen

- Maß, Dichte, Erwartungswert, Zufallsvariable, ...

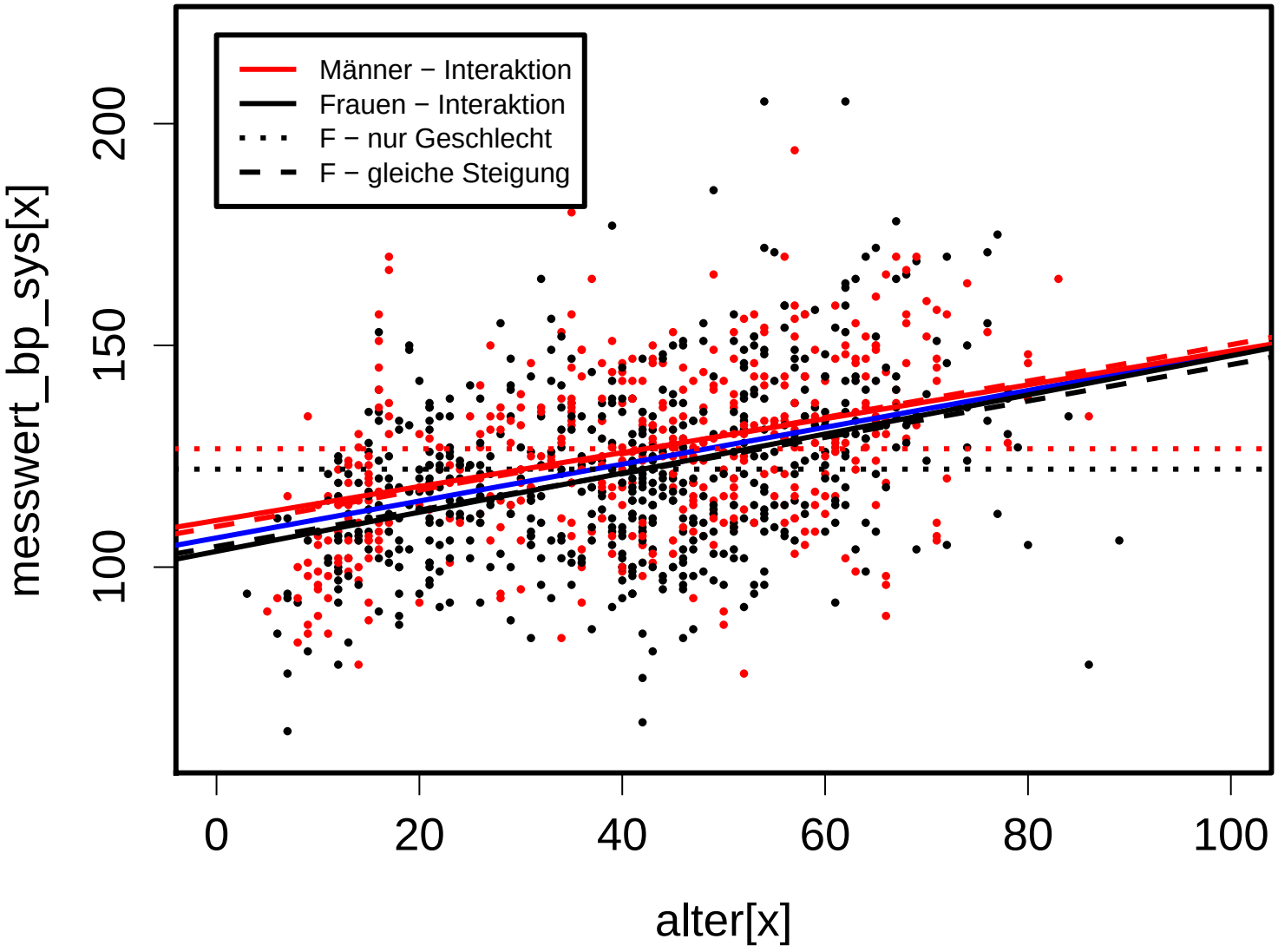
$$\mathbb{P}(\boldsymbol{X} \in S) = \int_S \frac{\mathrm{d}\,\mathbb{P} \circ \boldsymbol{X}^{-1}}{\mathrm{d}\,\nu}\,\mathrm{d}\,\nu = \int_S p\,\mathrm{d}\,\nu$$

$$p(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{p(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})}$$

$$\mathbb{E}[\phi(\boldsymbol{X})] = \int_{\mathcal{X}} \phi\,\mathrm{d}\,\nu$$



Regression


Männer − Interaktion
Frauen − Interaktion
F − nur Geschlecht
F − gleiche Steigung
messwert_bp_sys[x]
100
150
200
0
20
40
60
80
100
alter[x]

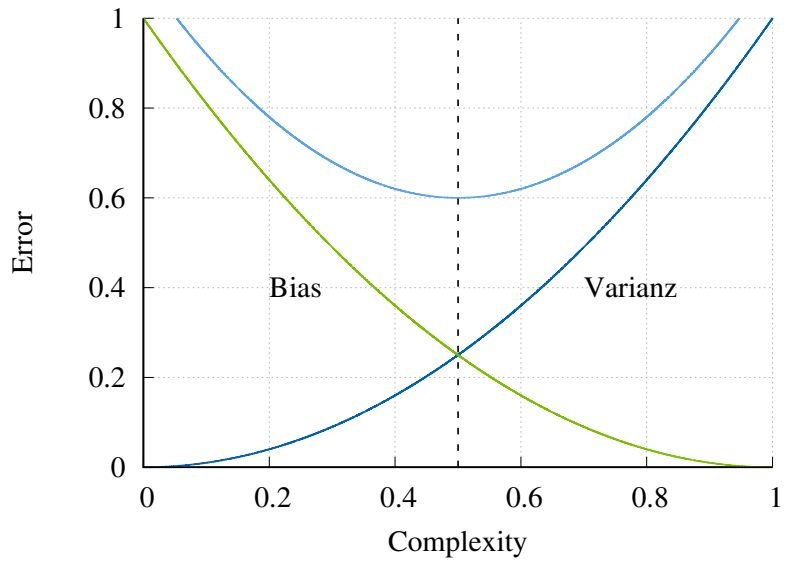
Maschinelles Lernen

- Optimierung, Modell, Verlustfunktion, Overfitting, . . .

$$\mathbb{E}[(y - \hat{f}(\boldsymbol{x}))^2] = \mathbb{B}[\hat{f}(\boldsymbol{x})^2] + \mathbb{V}[\hat{f}(\boldsymbol{x})] + \sigma^2$$

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_t \nabla \ell(\boldsymbol{\beta}^t; \mathcal{D})$$

$$\hat{f} = \min_{f \in \mathcal{F}} \ell(f; \mathcal{D})$$



Frequent Set Mining

- Datenbanken, Häufige Mengen, A-Priori, FP-Trees, ...

```
SELECT * FROM transactions WHERE ...
```

$$p(A, B) \leq \min\{p(A), p(B)\}$$

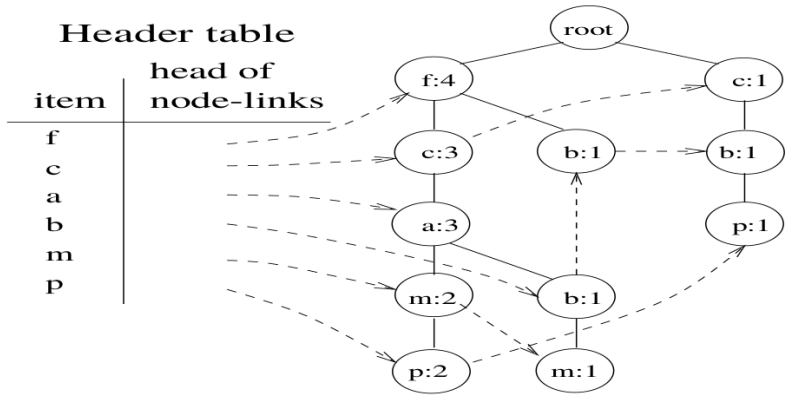


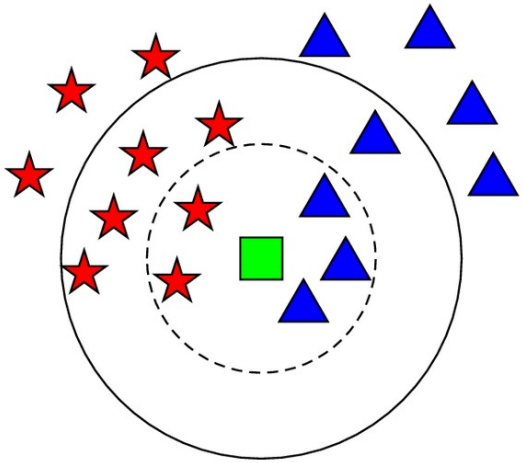
Figure 1: The FP-tree in Example 1.

Klassifikation

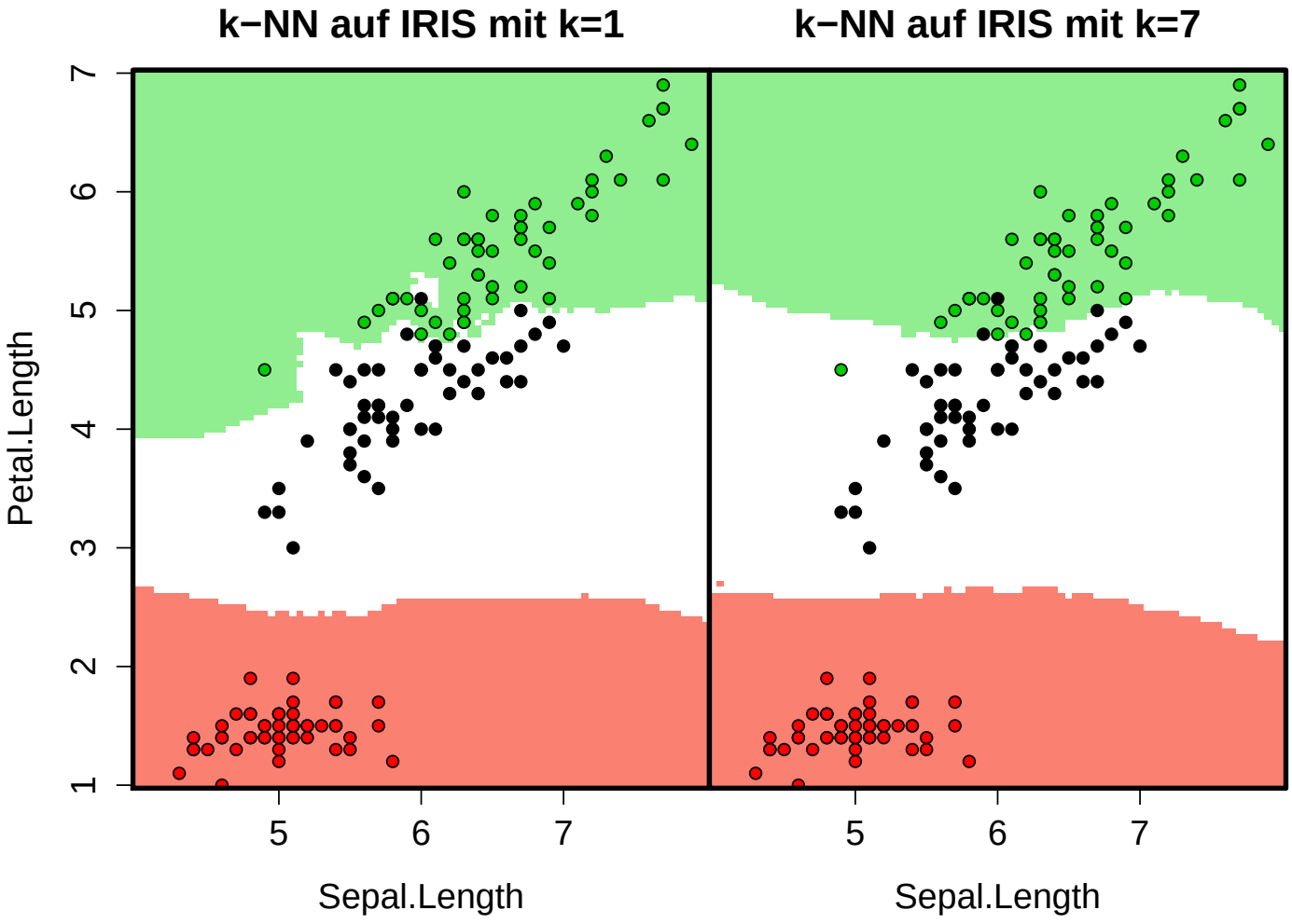
- Logistische Regression, k-NN, Distanzmaße, ...

$$p(y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle))}$$

$$\text{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{\sum_{i=1}^{d} (\boldsymbol{x}_i - \boldsymbol{y}_i)^2}$$



Klassifikation



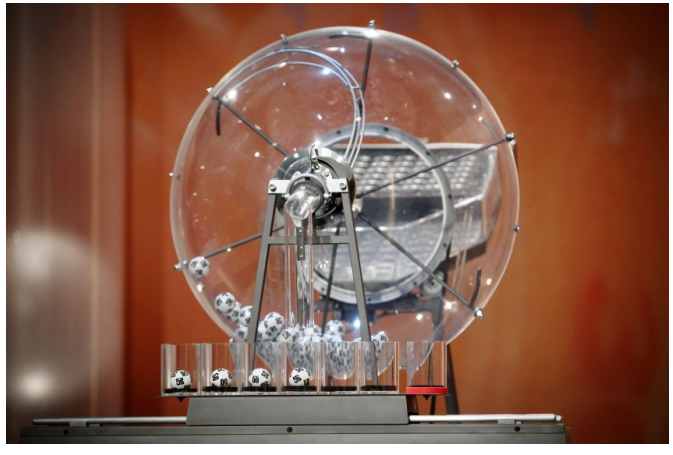
Modelle und Sampling

- Modellselektion, Sampling, Klassifikationsgüte, ...

$$\text{BIC} = d\log(N) - 2\ell(\boldsymbol{\beta}^*; \mathcal{D})$$

$$F_1 = \frac{2 \times \text{PREC} \times \text{REC}}{\text{PREC} + \text{REC}}$$

$$\boldsymbol{x} \sim \mathbb{P}$$



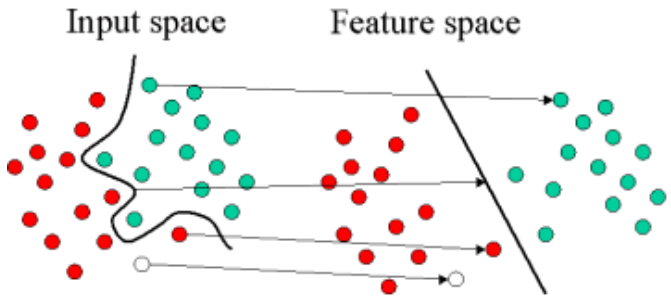
Stützvektormethode

- Hyperebene, Hinge-Loss, Merkmalsraum, Kernel, ...

$$f(\boldsymbol{x}) = b + \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle$$

$$K_{\text{Gauss}}(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle = \exp\left(-\frac{1}{\gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right)$$

$$\ell_{\text{Hinge}}(y^*, \boldsymbol{x}, \boldsymbol{\beta}) = \max\{0, 1 - y^* \times (b + \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle)\}$$



Diskriminanzanalyse

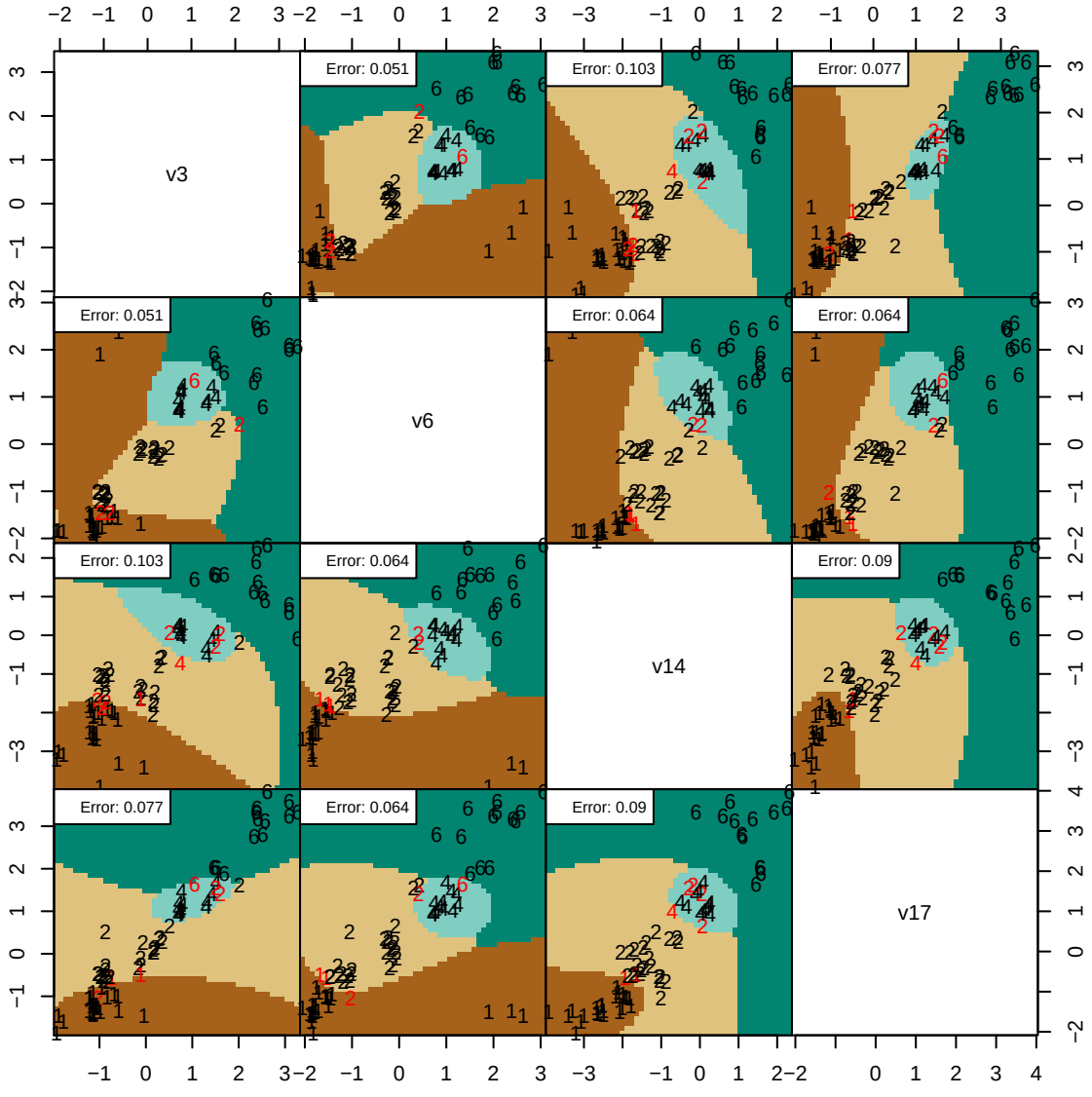
- Linear, Quadratic, Regularized, . . .

$$h_i^L(x) := (\Sigma^{-1}\mu_i)'x - 0.5\mu_i'\Sigma^{-1}\mu_i + \ln(\pi_i)$$

$$h_i^Q(x) := -0.5(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln(\pi_i) - 0.5\ln(\det(\Sigma_i))$$

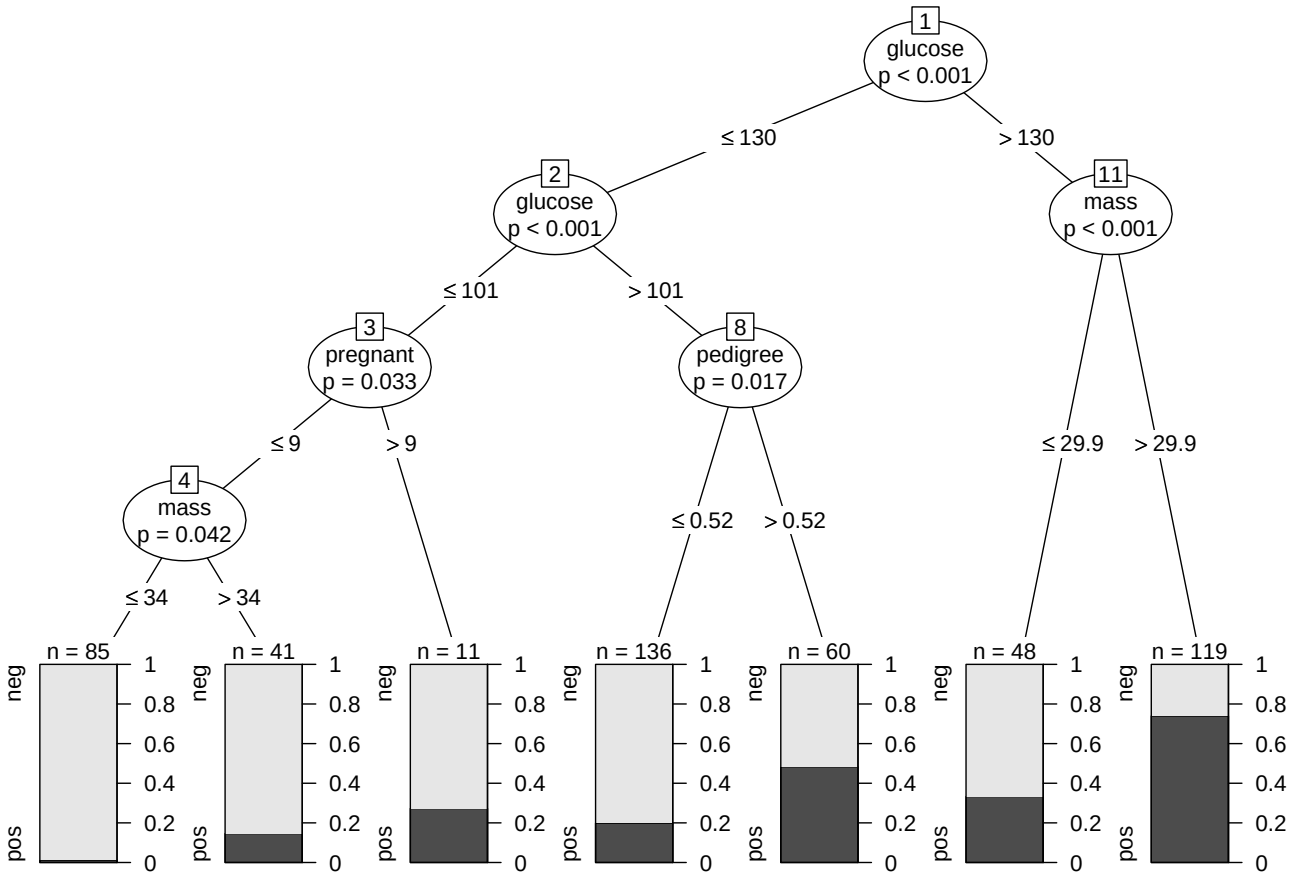
$$\hat{\Sigma}_i(\delta,\lambda) := (1-\lambda)\hat{\Sigma}_i(\delta) + \frac{\lambda \cdot tr[\hat{\Sigma}_i(\delta)]}{p} I$$

Diskriminanzanalyse



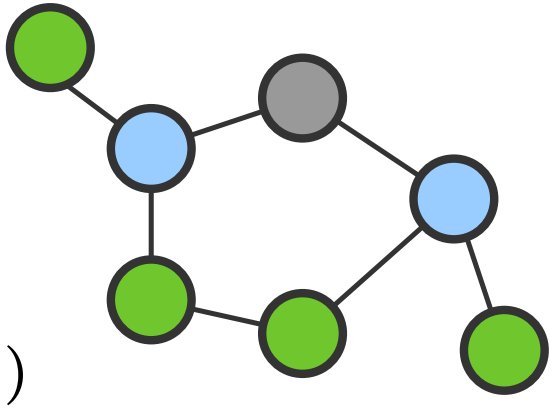
Ensembles

- Bäume, Bagging, Boosting, Radom Forests . . .



Graphische Modelle

- Exponentialfamilien, Markov Random Fields, Belief-Propagation, ...



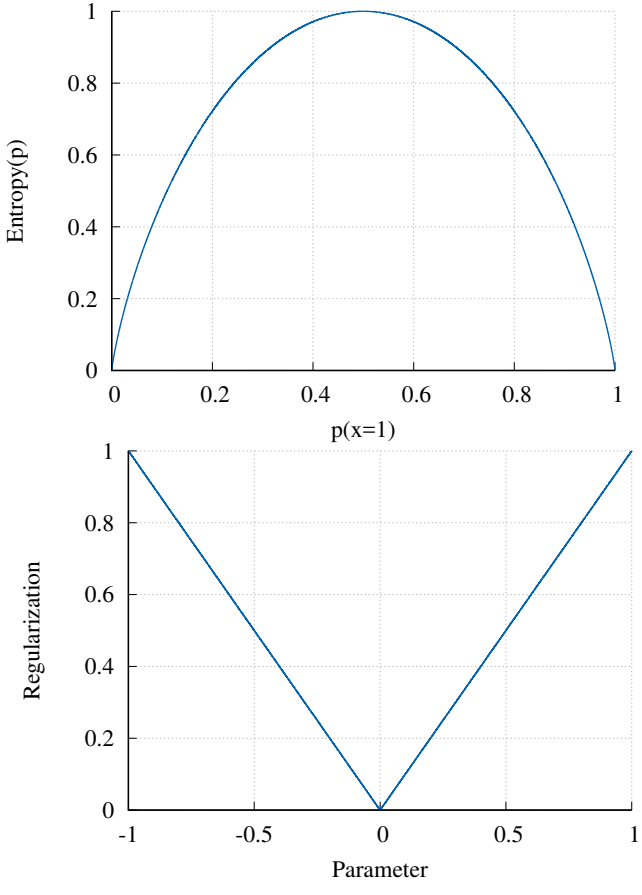
$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta}))$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z} \prod_{v \in V} \psi_v(\boldsymbol{x}_v) \prod_{(v,u) \in E} \psi_{vu}(\boldsymbol{x}_v, \boldsymbol{x}_u)$$

$$\boldsymbol{m}_{u \to v}(x) = \sum_{y \in \mathcal{X}_u} \psi_u(y) \psi_{uv}(x, y) \prod_{w \in \mathcal{N}_u \setminus \{v\}} \boldsymbol{m}_{w \to u}(y)$$

Strukturlernen

- Merkmalsselektion, Regularisierung, nicht-glatte Optimierung, . . .



$$I(v,u) = \sum_{x,y} p_{vu}(x,y) \log \frac{p_{vu}(x,y)}{p_v(x)p_u(y)}$$

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \ell(\boldsymbol{\beta}; \mathcal{D}) + \lambda R(\boldsymbol{\beta})$$

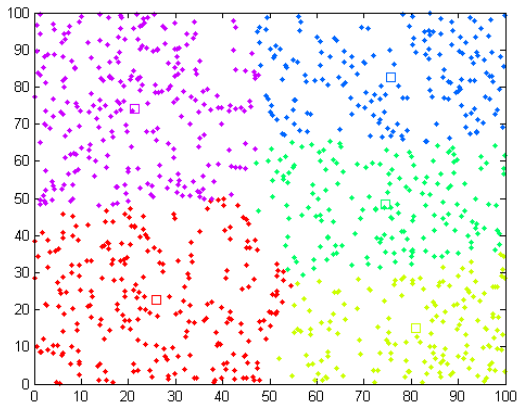
$$\text{prox}_{\lambda \|\cdot\|_1}(\boldsymbol{\beta})_i = \begin{cases} \boldsymbol{\beta}_i - \lambda & , \boldsymbol{\beta}_i \geq +\lambda \\ 0 & , |\boldsymbol{\beta}_i| < \lambda \\ \boldsymbol{\beta}_i + \lambda & , \boldsymbol{\beta}_i \leq -\lambda \end{cases}$$

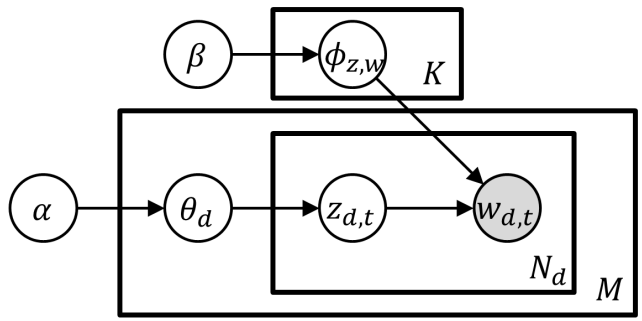
Clusteranalyse

- k-Means, Gaussian Mixture, Latent Dirichlet Allocation, ...

$$\min_{C \subseteq \mathbb{R}^d, |C|=k} \sum_{\boldsymbol{x} \in \mathcal{D}} \min_{\boldsymbol{c} \in C} \|\boldsymbol{x} - \boldsymbol{c}\|_2$$

$$p(\boldsymbol{w}_i^d \mid \boldsymbol{z}_i^d, \boldsymbol{\phi}) p(\boldsymbol{z}_i^d \mid \boldsymbol{\theta}^d) p(\boldsymbol{\theta}^d \mid \boldsymbol{\alpha}) p(\boldsymbol{\phi} \mid \boldsymbol{\beta})$$





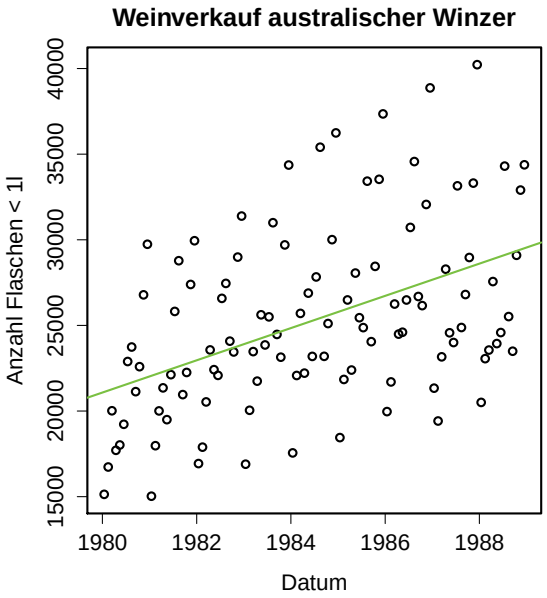
Zeitreihenanalyse

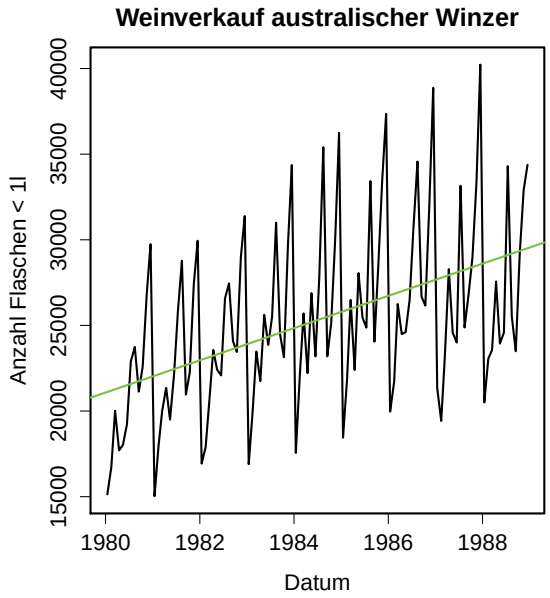
- Glättung, von Autoregression bis SARIMA, Schwingungen

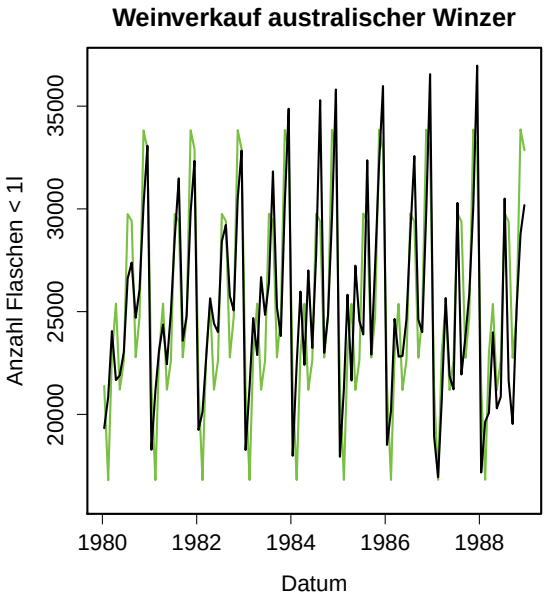
$$y_t = \beta_1 + \beta_2 y_{t-1} + \ldots + \beta_{p+1} y_{t-p} + \epsilon_t - \gamma_1 \epsilon_{t-1} - \ldots - \gamma_q \epsilon_{t-q}$$

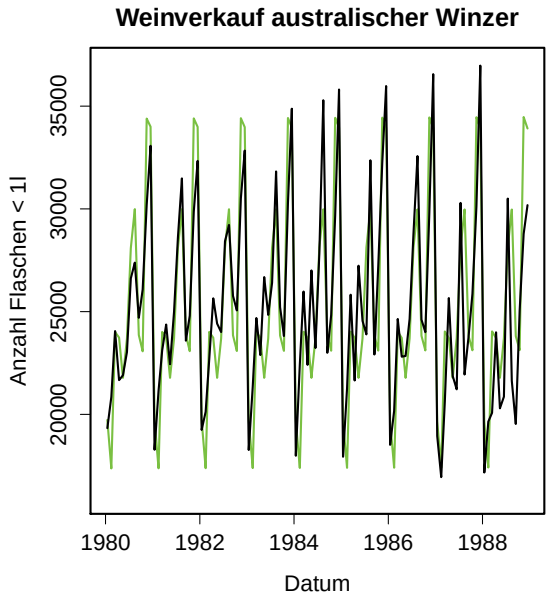
$$y_t = \beta_1 + \sum_{k=1}^{K} (\beta_{2k} \cos(2\pi\lambda_k t) + \beta_{2k+1} \sin(2\pi\lambda_k t)) + \epsilon_t$$

Zeitreihenanalyse





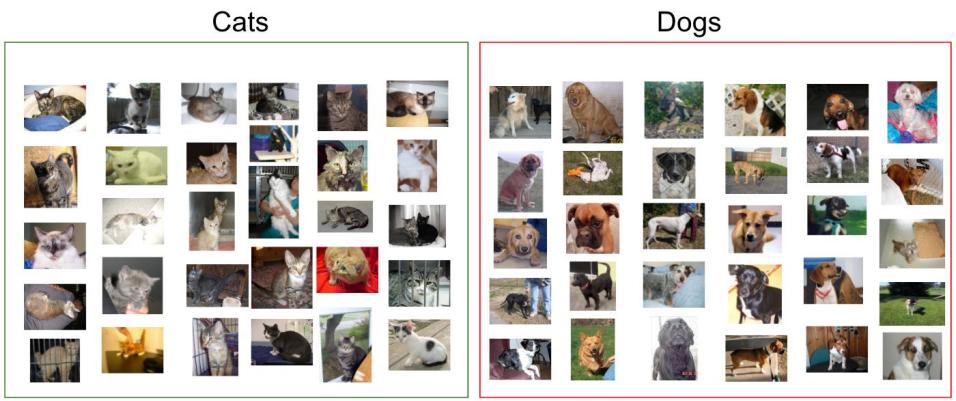




Künstliche Neuronale Netze

- Versteckte Schichten, Back-Propagation, LRU, Embeddings, . . .

$$\frac{\partial \operatorname{Err}(\boldsymbol{x}, \boldsymbol{y}^*, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{l,i,j}} = \frac{\partial \operatorname{Err}(\boldsymbol{x}, \boldsymbol{y}^*, \boldsymbol{\beta})}{\partial \operatorname{Out}^\top} \frac{\partial \operatorname{Out}}{\partial \operatorname{Net}_n^\top} \cdots \frac{\partial \operatorname{Net}_{l+1}}{\partial \operatorname{Net}_l^\top} \left(\frac{\partial \operatorname{Net}_l}{\partial \boldsymbol{\beta}_{l,i,j}} \right)^\top$$



Sample of cats & dogs images from Kaggle Dataset