Wissensentdeckung in Datenbanken

Modellklassen, Verlustfunktionen

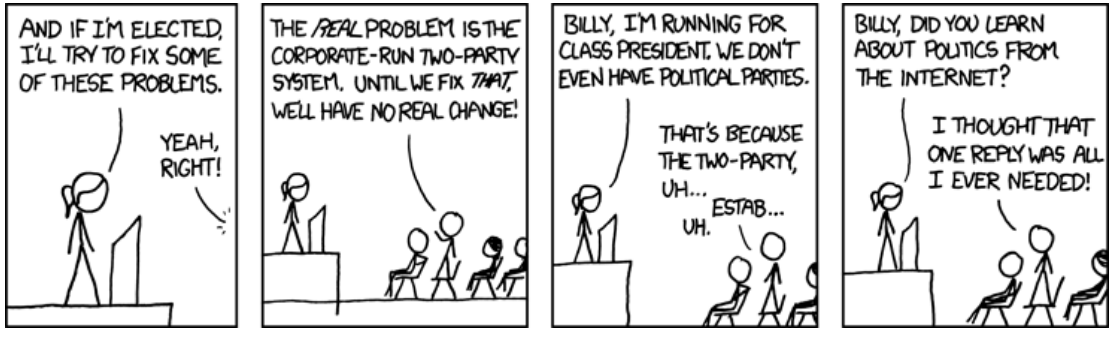
Nico Piatkowski und Uwe Ligges

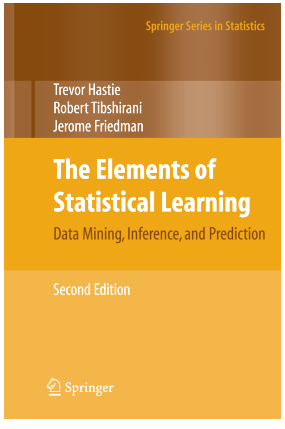
Informatik—Künstliche Intelligenz
Computergestützte Statistik
Technische Universität Dortmund

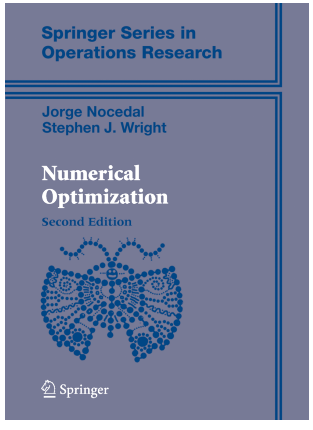
02.05.2017

Literatur

- Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. Springer Series in Statistics. Springer. 2009
  - Im Moodle zum Download verfügbar.
- Jorge Nocedal. Stephen Wright. Numerical Optimization. 2nd edition. Springer Series in Operations Research and Financial Engineering. Springer-Verlag New York. 2006
  - In der Zentralbibliothek verfügbar.







Daten—und dann?

- Personendaten
- Medizinische Daten
- Konto- und Zahlungsdaten
- Verbindungsdaten
- Soziale Netzwerke

Realisierung von Zufallsvektoren X

$$X = \begin{pmatrix} \text{Alter} \\ \text{Geschlecht} \\ \text{Krankheitstage} \\ \text{Medikation} \\ \text{Kontostand} \\ \text{Kredite} \\ \text{Webseiten} \\ \dots \end{pmatrix}$$

Begriffe

- Daten $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), \ldots, (\boldsymbol{x}^N, \boldsymbol{y}^N)\}$
  - Realisierungen von Zufallsvariable $\boldsymbol{X}$ mit n-dimensionaler Domäne $\mathcal{X}$
  - Multimenge; $\#_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{N}$
  - Messfehler, Rauschen $\boldsymbol{\epsilon} \sim \mathbb{P}$
  - Fehlende Werte; $\boldsymbol{x}_i \in \mathcal{X}_i \cup \{?\}, 1 \leq i \leq n$
- Modell f aus Modellklasse $\mathcal{M}$
  - Funktionen $\mathcal{M} \subseteq \mathcal{F}$
  - Koeffizienten / Parameter $\mathcal{M} \subseteq \mathbb{R}^d$
  - Datenpunkte $\mathcal{M} \subseteq \mathcal{D}$, $\mathcal{M} \subseteq \mathcal{X} \times \mathcal{Y}$
- Verlustfunktion (Güte) $\ell : (f; \mathcal{D}) \mapsto \mathbb{R}$

Begriffe

- Daten $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), \ldots, (\boldsymbol{x}^N, \boldsymbol{y}^N)\}$
  - Realisierungen von Zufallsvariable $\boldsymbol{X}$ mit n-dimensionaler Domäne $\mathcal{X}$
  - Multimenge; $\#_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{N}$
  - Messfehler, Rauschen $\boldsymbol{\epsilon} \sim \mathbb{P}$
  - Fehlende Werte; $\boldsymbol{x}_i \in \mathcal{X}_i \cup \{?\}, 1 \leq i \leq n$
- Modell f aus Modellklasse $\mathcal{M}$
  - Funktionen $\mathcal{M} \subseteq \mathcal{F}$
  - Koeffizienten / Parameter $\mathcal{M} \subseteq \mathbb{R}^d$
  - Datenpunkte $\mathcal{M} \subseteq \mathcal{D}$, $\mathcal{M} \subseteq \mathcal{X} \times \mathcal{Y}$
- Verlustfunktion (Güte) $\ell : (f; \mathcal{D}) \mapsto \mathbb{R}$

Begriffe

- Daten $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), \ldots, (\boldsymbol{x}^N, \boldsymbol{y}^N)\}$
  - Realisierungen von Zufallsvariable $\boldsymbol{X}$ mit n-dimensionaler Domäne $\mathcal{X}$
  - Multimenge; $\#_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{N}$
  - Messfehler, Rauschen $\boldsymbol{\epsilon} \sim \mathbb{P}$
  - Fehlende Werte; $\boldsymbol{x}_i \in \mathcal{X}_i \cup \{?\}, 1 \leq i \leq n$
- Modell f aus Modellklasse $\mathcal{M}$
  - Funktionen $\mathcal{M} \subseteq \mathcal{F}$
  - Koeffizienten / Parameter $\mathcal{M} \subseteq \mathbb{R}^d$
  - Datenpunkte $\mathcal{M} \subseteq \mathcal{D}$, $\mathcal{M} \subseteq \mathcal{X} \times \mathcal{Y}$
- Verlustfunktion (Güte) $\ell : (f; \mathcal{D}) \mapsto \mathbb{R}$

Allgemeines Vorgehen

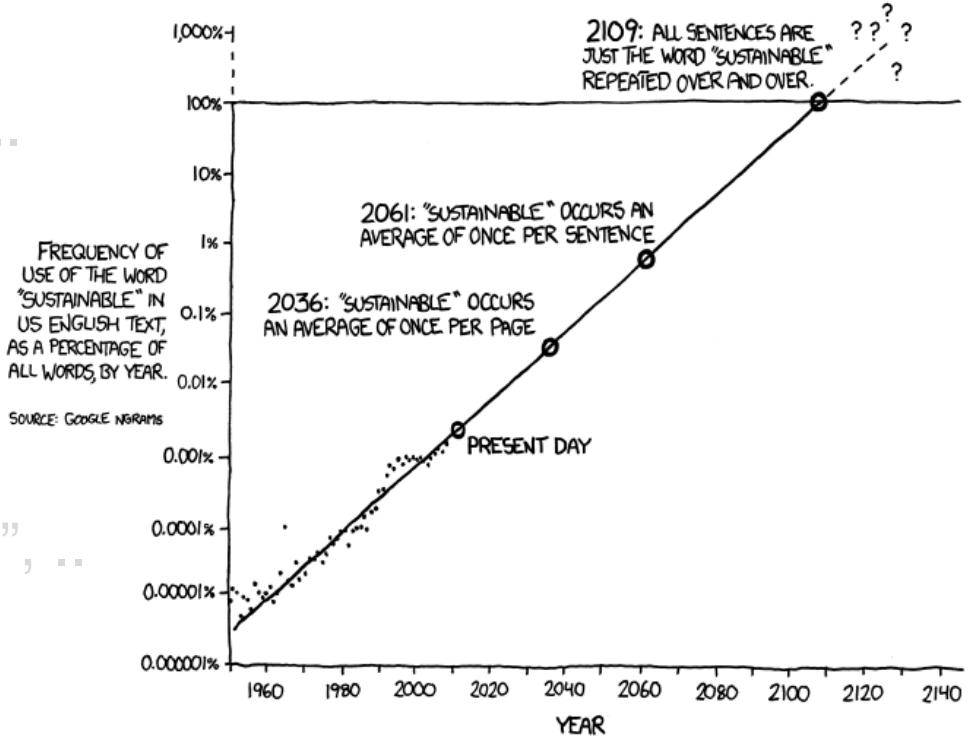
- Daten $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), \ldots, (\boldsymbol{x}^N, \boldsymbol{y}^N)\}$
- Datenpunkt $\boldsymbol{x} \in \mathcal{X}$, Label (Klasse) $\boldsymbol{y} \in \mathcal{Y}$
- Modelle $\mathcal{M}$

Modell “lernen”, Datenanalyse, ..

$$f^* = \arg\min_{f \in \mathcal{M}} \ell(f; \mathcal{D})$$

Modellanwendung, “Vorhersage”, ..

$$\hat{\boldsymbol{y}} = f^*(\boldsymbol{x})$$



Allgemeines Vorgehen

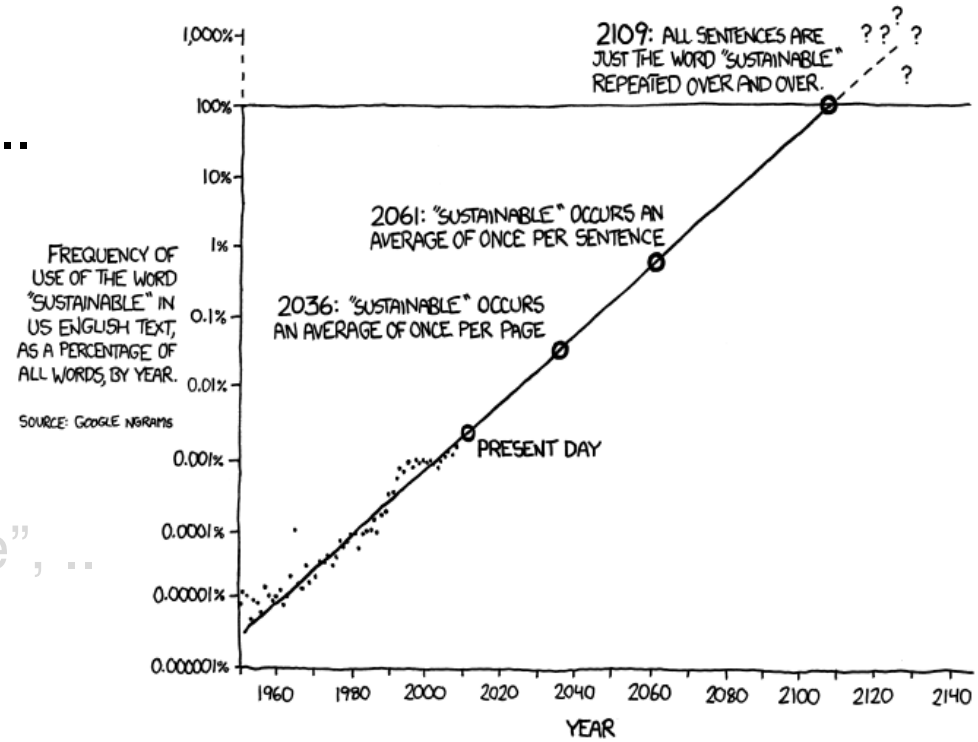
- Daten $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), \ldots, (\boldsymbol{x}^N, \boldsymbol{y}^N)\}$
- Datenpunkt $\boldsymbol{x} \in \mathcal{X}$, Label (Klasse) $\boldsymbol{y} \in \mathcal{Y}$
- Modelle $\mathcal{M}$

Modell “lernen”, Datenanalyse, ..

$$f^* = \arg\min_{f \in \mathcal{M}} \ell(f; \mathcal{D})$$

Modellanwendung, “Vorhersage”, ..

$$\hat{\boldsymbol{y}} = f^*(\boldsymbol{x})$$



Allgemeines Vorgehen

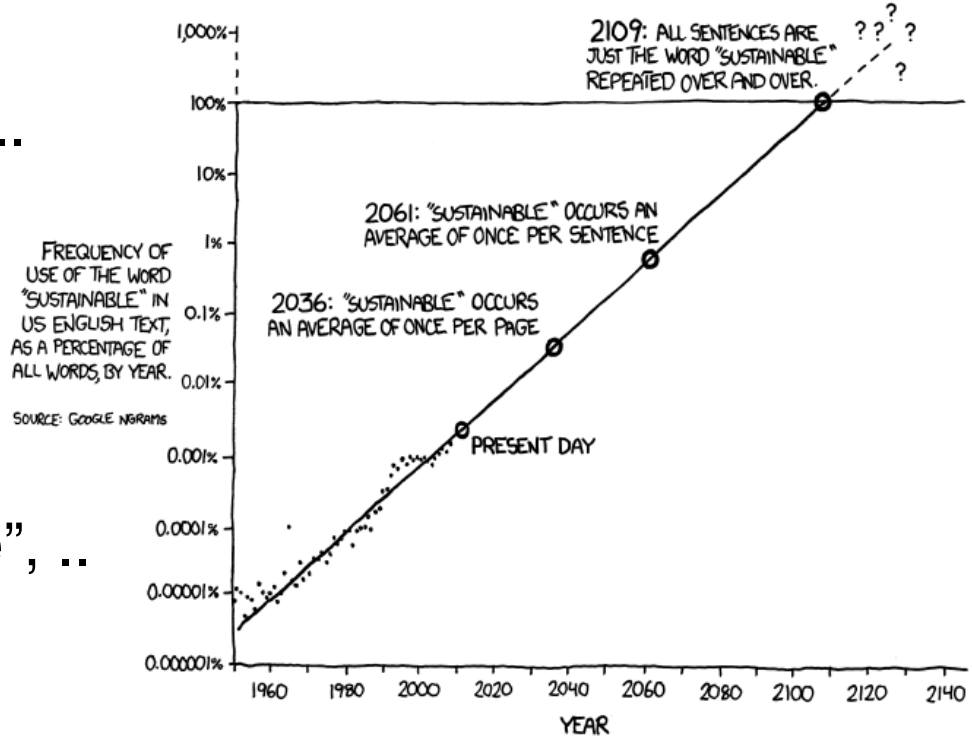
- Daten $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), \ldots, (\boldsymbol{x}^N, \boldsymbol{y}^N)\}$
- Datenpunkt $\boldsymbol{x} \in \mathcal{X}$, Label (Klasse) $\boldsymbol{y} \in \mathcal{Y}$
- Modelle $\mathcal{M}$

Modell “lernen”, Datenanalyse, ..

$$f^* = \arg\min_{f \in \mathcal{M}} \ell(f; \mathcal{D})$$

Modellanwendung, “Vorhersage”, ..

$$\hat{\boldsymbol{y}} = f^*(\boldsymbol{x})$$

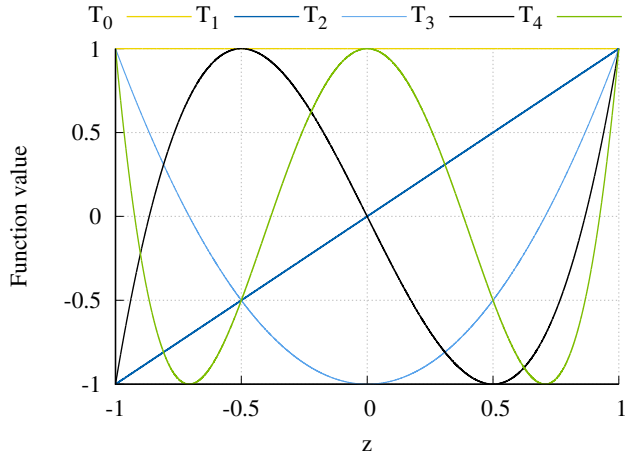


Modellklassen $\mathcal{M}$

$$f : \mathcal{X} \to \mathcal{Y}$$

- Linear

$$f(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$$



- Polynom

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i \boldsymbol{x}_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{i,j} \boldsymbol{x}_i \boldsymbol{x}_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \beta_{i,j,k} \boldsymbol{x}_i \boldsymbol{x}_j \boldsymbol{x}_k + \dots$$

- Trigonometrisch/Periodisch

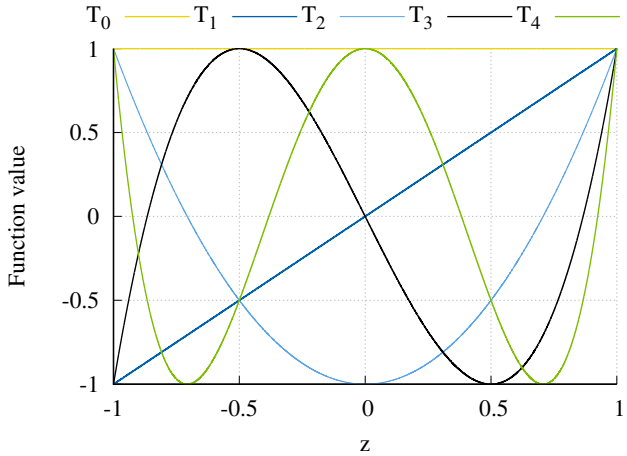
$$f(\boldsymbol{x}) = \sum_{i=0}^{k} \boldsymbol{\theta}_i \cos(i \arccos(\langle \boldsymbol{\beta}^i, \boldsymbol{x} \rangle))$$

Modellklassen $\mathcal{M}$

$$f : \mathcal{X} \to \mathcal{Y}$$

- Linear

$$f(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$$



- Polynom

$$f(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{x}_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\beta}_{i,j} \boldsymbol{x}_i \boldsymbol{x}_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \boldsymbol{\beta}_{i,j,k} \boldsymbol{x}_i \boldsymbol{x}_j \boldsymbol{x}_k + \dots$$

- Trigonometrisch/Periodisch

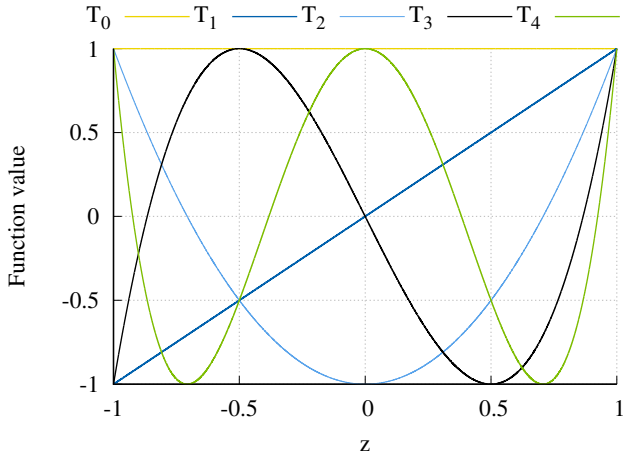
$$f(\boldsymbol{x}) = \sum_{i=0}^{k} \boldsymbol{\theta}_i \cos(i \arccos(\langle \boldsymbol{\beta}^i, \boldsymbol{x} \rangle))$$

Modellklassen $\mathcal{M}$

$$f : \mathcal{X} \to \mathcal{Y}$$

- Linear

$$f(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$$



- Polynom

$$f(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{x}_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\beta}_{i,j} \boldsymbol{x}_i \boldsymbol{x}_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \boldsymbol{\beta}_{i,j,k} \boldsymbol{x}_i \boldsymbol{x}_j \boldsymbol{x}_k + \ldots$$

- Trigonometrisch/Periodisch

$$f(\boldsymbol{x}) = \sum_{i=0}^{k} \boldsymbol{\theta}_i \cos(i \arccos(\langle \boldsymbol{\beta}^i, \boldsymbol{x} \rangle))$$

Modellklassen $\mathcal{M}$ (II)

- Probabilistisch/Exponentialfamilie/Bayesianisch

$$p_{\text{Gauss}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$$p_{\text{Posterior}}(\boldsymbol{x}) = p_{\text{Likelihood}}(\boldsymbol{x}) p_{\text{Prior}_1}(\alpha) p_{\text{Prior}_2}(\gamma) \dots$$

- Unstetig/Piecewise/Thresholding

$$f_v(\boldsymbol{x}) = \begin{cases} f_{\text{Left}(v)}(\boldsymbol{x}) & g(\boldsymbol{x}) \geq \rho_v \\ f_{\text{Right}(v)}(\boldsymbol{x}) & g(\boldsymbol{x}) < \rho_v \end{cases}$$

Modellklassen $\mathcal{M}$ (II)

- Probabilistisch/Exponentialfamilie/Bayesianisch

$$p_{\text{Gauss}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$$p_{\text{Posterior}}(\boldsymbol{x}) = p_{\text{Likelihood}}(\boldsymbol{x}) p_{\text{Prior}_1}(\alpha) p_{\text{Prior}_2}(\gamma) \dots$$

- Unstetig/Piecewise/Thresholding

$$f_v(\boldsymbol{x}) = \begin{cases} f_{\text{Left}(v)}(\boldsymbol{x}) & g(\boldsymbol{x}) \geq \rho_v \\ f_{\text{Right}(v)}(\boldsymbol{x}) & g(\boldsymbol{x}) < \rho_v \end{cases}$$

Beispiel: Fehlerwahrscheinlichkeit

- $\mathcal{Y} = \{c_1, c_2, \ldots, c_k\}$, Modell f
- Wahrscheinlichkeitsdichte falscher Vorhersagen:

$$p(f(\boldsymbol{X}) \neq \boldsymbol{Y}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}[\mathbb{1}_{\{f(\boldsymbol{X})\neq\boldsymbol{Y}\}}] = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\mathbb{1}_{\{f(\boldsymbol{X})\neq\boldsymbol{Y}\}}]$$
$$= \mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{k} \mathbb{1}_{\{f(\boldsymbol{X})\neq c_i\}} \mathbb{P}(\boldsymbol{Y} = c_i \mid \boldsymbol{X})\right]$$

- Wahl von f, so dass innere Summe minimiert wird..

$$f(\boldsymbol{x}) = \arg\min_{j=1}^{k} \sum_{i=1}^{k} \mathbb{1}_{\{c_j\neq c_i\}} \mathbb{P}(\boldsymbol{Y} = c_i \mid \boldsymbol{X} = \boldsymbol{x})$$
$$= \arg\min_{j=1}^{k} (1 - \mathbb{P}(\boldsymbol{Y} = c_j \mid \boldsymbol{X} = \boldsymbol{x}))$$
$$= \arg\max_{j=1}^{k} \mathbb{P}(\boldsymbol{Y} = c_j \mid \boldsymbol{X} = \boldsymbol{x})$$

Beispiel: Fehlerwahrscheinlichkeit

- $\mathcal{Y} = \{c_1, c_2, \ldots, c_k\}$, Modell f
- Wahrscheinlichkeitsdichte falscher Vorhersagen:

$$\begin{aligned} p(f(\boldsymbol{X}) \neq \boldsymbol{Y}) &= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\mathbb{1}_{\{f(\boldsymbol{X})\neq \boldsymbol{Y}\}}\right] = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\mathbb{1}_{\{f(\boldsymbol{X})\neq \boldsymbol{Y}\}}\right] \\ &= \mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{k} \mathbb{1}_{\{f(\boldsymbol{X})\neq c_i\}} \mathbb{P}(\boldsymbol{Y} = c_i \mid \boldsymbol{X})\right] \end{aligned}$$

- Wahl von f, so dass innere Summe minimiert wird..

$$\begin{aligned} f(\boldsymbol{x}) &= \arg\min_{j=1}^{k} \sum_{i=1}^{k} \mathbb{1}_{\{c_j \neq c_i\}} \mathbb{P}(\boldsymbol{Y} = c_i \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \arg\min_{j=1}^{k} (1 - \mathbb{P}(\boldsymbol{Y} = c_j \mid \boldsymbol{X} = \boldsymbol{x})) \\ &= \arg\max_{j=1}^{k} \mathbb{P}(\boldsymbol{Y} = c_j \mid \boldsymbol{X} = \boldsymbol{x}) \end{aligned}$$

Verlustfunktionen ℓ

- Absoluter Fehler/SSE/MSE/RMSE

$$\mathrm{Err}(f;\mathcal{D}) = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} |y - f(\boldsymbol{x})|$$

$$\mathrm{RMSE}(f;\mathcal{D}) = \sqrt{\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} (y - f(\boldsymbol{x}))^2}$$

- Hinge Loss

$$\mathrm{Hinge}(f;\mathcal{D}) = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \max\{0, 1 - yf(\boldsymbol{x})\}$$

Verlustfunktionen ℓ

- Absoluter Fehler/SSE/MSE/RMSE

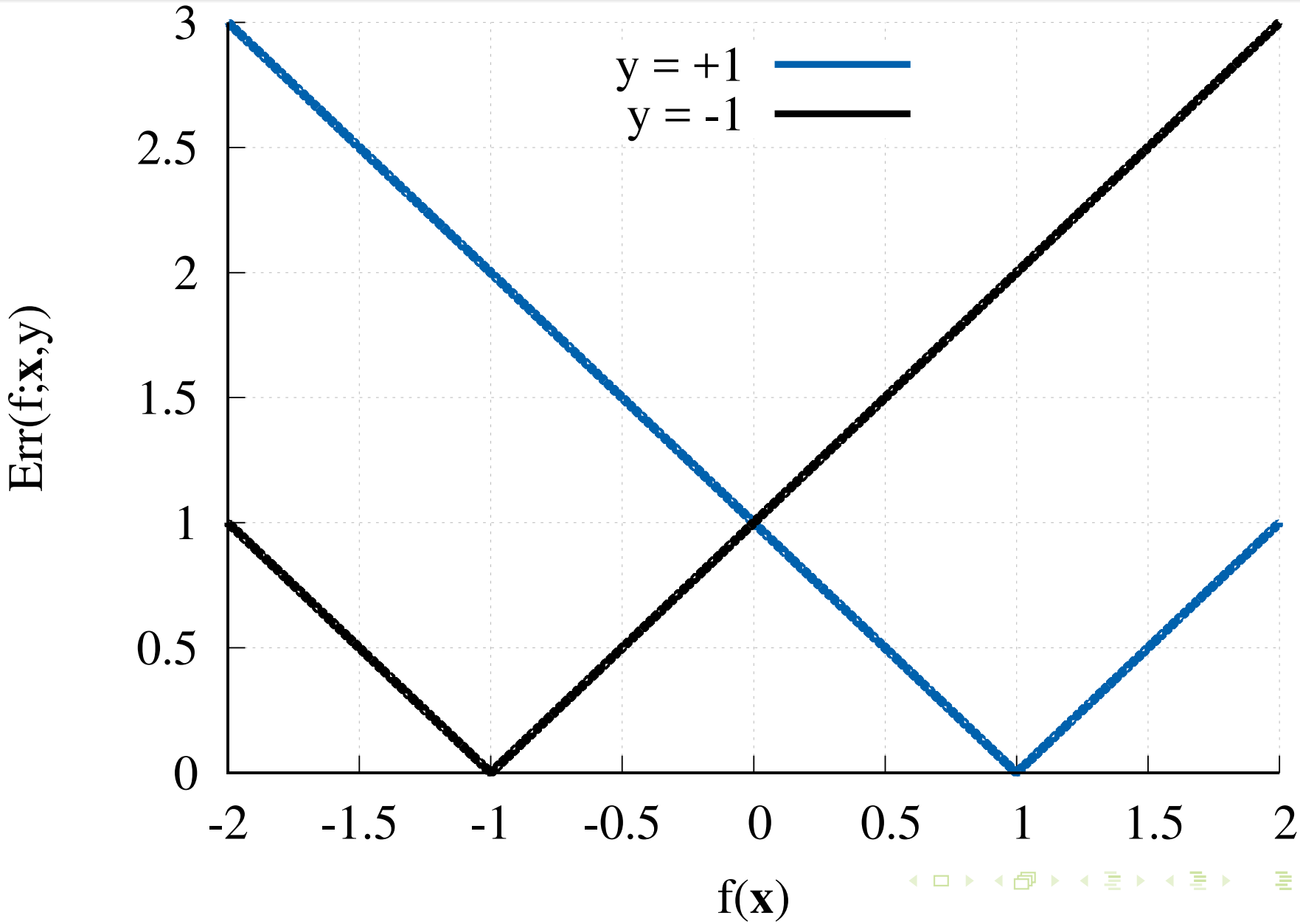
$$\mathrm{Err}(f;\mathcal{D}) = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} |y - f(\boldsymbol{x})|$$

$$\mathrm{RMSE}(f;\mathcal{D}) = \sqrt{\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} (y - f(\boldsymbol{x}))^2}$$

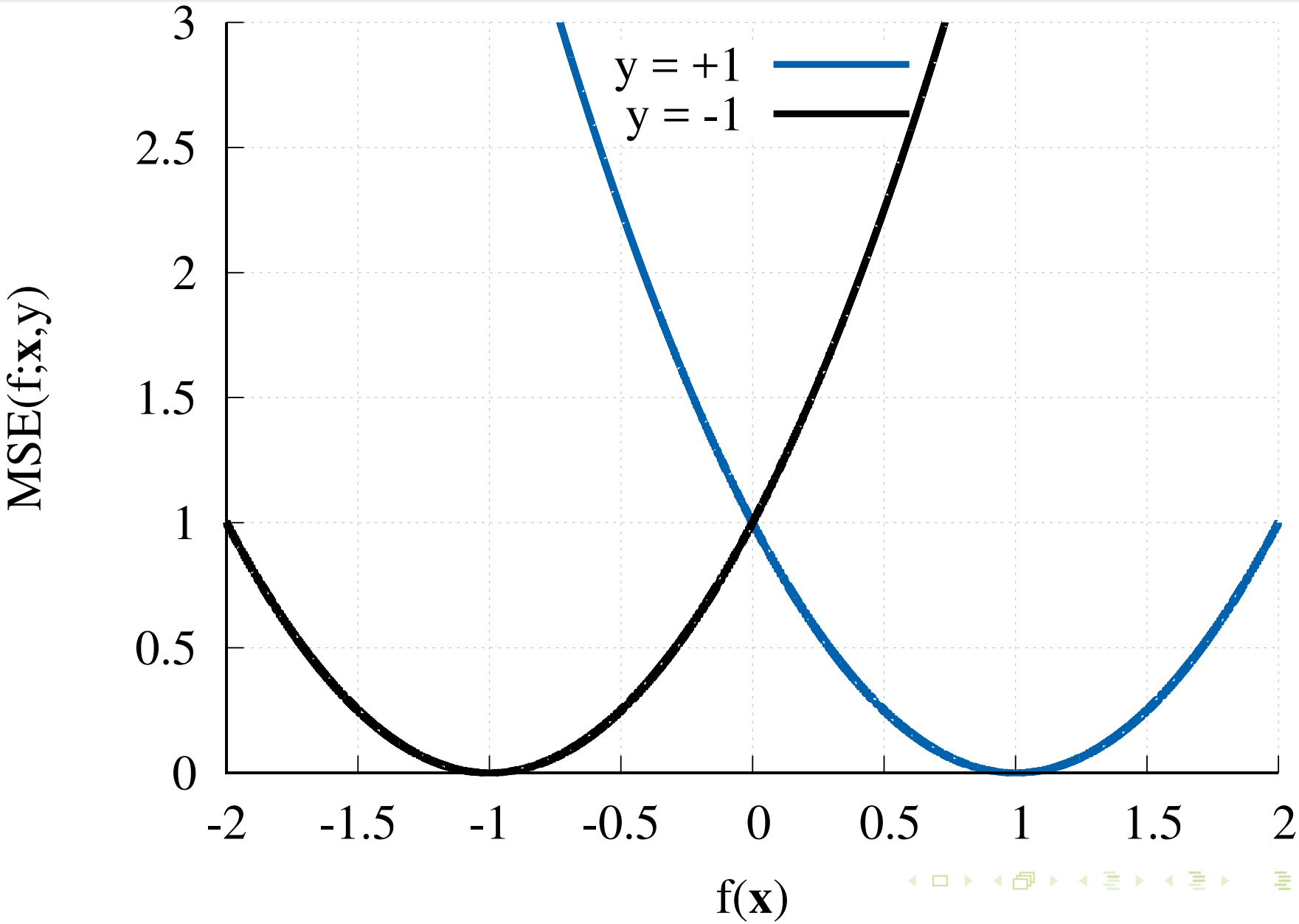
- Hinge Loss

$$\mathrm{Hinge}(f;\mathcal{D}) = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \max\{0, 1 - yf(\boldsymbol{x})\}$$

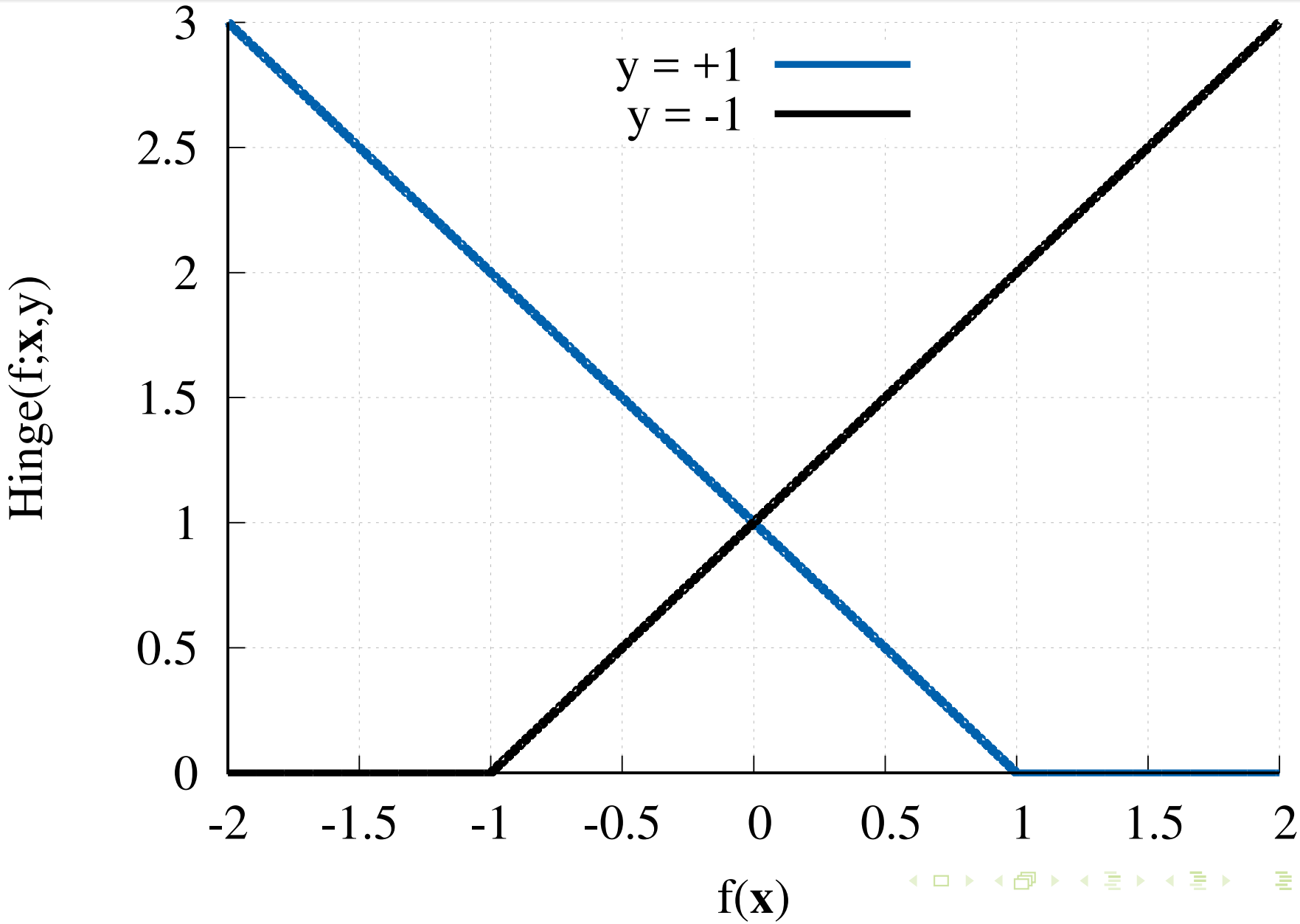
Absoluter Fehler, $|y - f(x)|$



Quadratischer Fehler, $(y - f(x))^2$



Hinge Fehler, $\max\{0, 1 - y - f(\boldsymbol{x})\}$



Verlustfunktionen ℓ (II) - Likelihood Varianten

- Likelihood

$$\mathcal{L}(p; \mathcal{D}) = \prod_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} p(\boldsymbol{x}, \boldsymbol{y})$$

- Log-Likelihood

$$\log \mathcal{L}(p; \mathcal{D}) = \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \log p(\boldsymbol{x}, \boldsymbol{y})$$

- Neg. Avg. Log-Likelihood

$$\ell(p; \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \log p(\boldsymbol{x}, \boldsymbol{y})$$

Verlustfunktionen ℓ (II) - Likelihood Varianten

- Likelihood

$$\mathcal{L}(p; \mathcal{D}) = \prod_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} p(\boldsymbol{x}, \boldsymbol{y})$$

- Log-Likelihood

$$\log \mathcal{L}(p; \mathcal{D}) = \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \log p(\boldsymbol{x}, \boldsymbol{y})$$

- Neg. Avg. Log-Likelihood

$$\ell(p; \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \log p(\boldsymbol{x}, \boldsymbol{y})$$

Verlustfunktionen ℓ (II) - Likelihood Varianten

- Likelihood

$$\mathcal{L}(p; \mathcal{D}) = \prod_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} p(\boldsymbol{x}, \boldsymbol{y})$$

- Log-Likelihood

$$\log \mathcal{L}(p; \mathcal{D}) = \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \log p(\boldsymbol{x}, \boldsymbol{y})$$

- Neg. Avg. Log-Likelihood

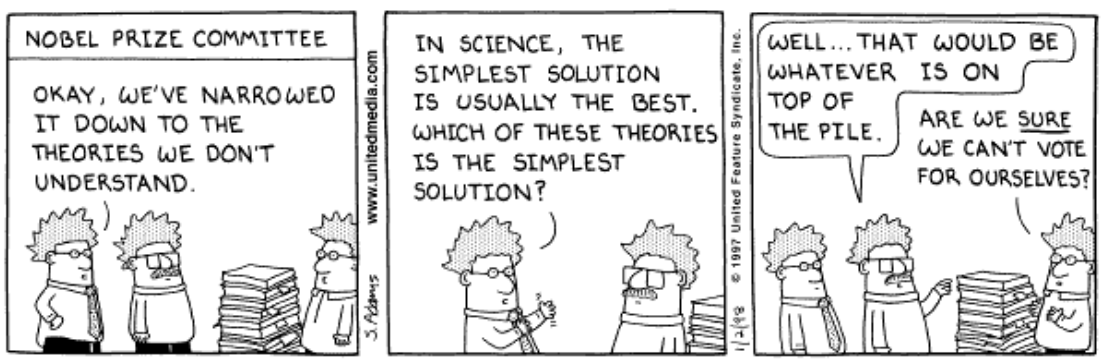
$$\ell(p; \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \log p(\boldsymbol{x}, \boldsymbol{y})$$

Verlustfunktionen ℓ (III) - MDL

- Minimum Description Length (MDL)
  - Formalisierung von Ockhams Rasiermesser

$$\min_{C \in \mathcal{M}} L(C) + L(\mathcal{D} \mid C)$$

- Intuition:
  - Nur wenige Objekte können kurze Codes haben
  - Nur wenige Objekte können hohe Wahrscheinlichkeit haben
- Formal: Alphabet $\mathcal{A} = \{1, 2, \ldots, m\}$, Codierung C, Codelängen $L_C(1), L_C(2), \ldots, L_C(m)$



$$\Leftrightarrow \sum_{a \in \mathcal{A}} 2^{-L_C(a)} \leq 1$$

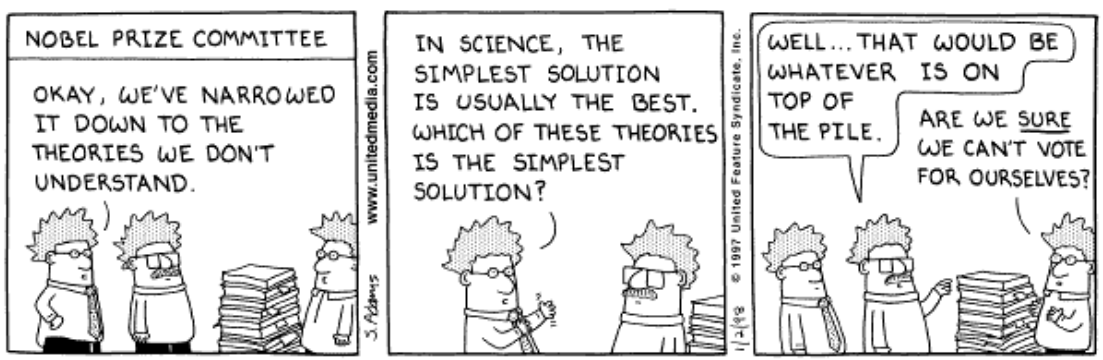
Kraft, 1949

Verlustfunktionen ℓ (III) - MDL

- Minimum Description Length (MDL)
  - Formalisierung von Ockhams Rasiermesser

$$\min_{C \in \mathcal{M}} L(C) + L(\mathcal{D} \mid C)$$

- Intuition:
  - Nur wenige Objekte können kurze Codes haben
  - Nur wenige Objekte können hohe Wahrscheinlichkeit haben
- Formal: Alphabet $\mathcal{A} = \{1, 2, \ldots, m\}$, Codierung C, Codelängen $L_C(1), L_C(2), \ldots, L_C(m)$



$$\Leftrightarrow \sum_{a \in \mathcal{A}} 2^{-L_C(a)} \leq 1$$

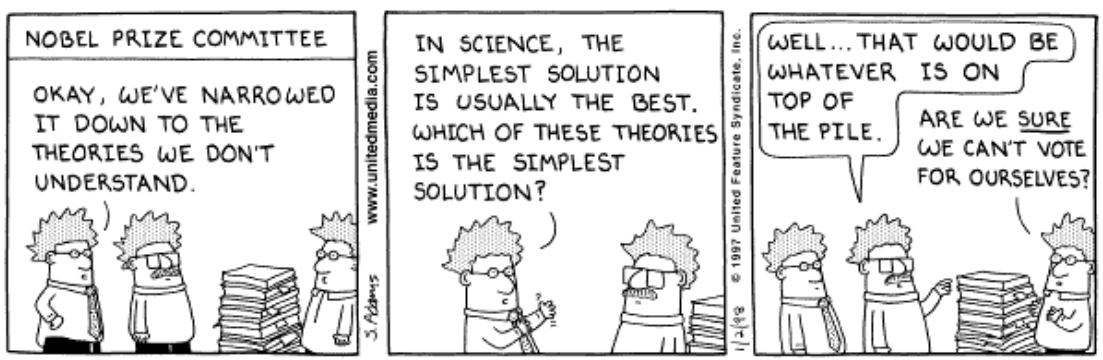
Kraft, 1949

Verlustfunktionen ℓ (III) - MDL

- Minimum Description Length (MDL)
  - Formalisierung von Ockhams Rasiermesser

$$\min_{C \in \mathcal{M}} L(C) + L(\mathcal{D} \mid C)$$

- Intuition:
  - Nur wenige Objekte können kurze Codes haben
  - Nur wenige Objekte können hohe Wahrscheinlichkeit haben
- Formal: Alphabet $\mathcal{A} = \{1, 2, \ldots, m\}$, Codierung C, Codelängen $L_C(1), L_C(2), \ldots, L_C(m)$



$$\Leftrightarrow \sum_{a \in \mathcal{A}} 2^{-L_C(a)} \leq 1$$

Kraft, 1949

Verlustfunktionen ℓ (IV) - Clustering

- Datensatz $\mathcal{D} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\}$ ohne Label
- Modell $\subset \mathbb{R}^n$
- k-Means (Intra-Cluster Varianz)

$$\mathrm{ICV}(k; \mathcal{D}) = \min_{C \subset \mathbb{R}^n, |C|=k} \sum_{\boldsymbol{x} \in \mathcal{D}} \min_{\boldsymbol{c} \in C} \|\boldsymbol{x} - \boldsymbol{c}\|_2^2$$

- Mixture Modelle (Expectation Maximization)
  - Example: Gaussian

$$\min_{\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{1 \le i \le k} \subset \mathbb{R}^n \times S^n_{++}} - \sum_{\boldsymbol{x} \in \mathcal{D}} \log \sum_{i=1}^{k} p_{\mathrm{Gauss}}(\boldsymbol{x} \mid i) p(i)$$

Verlustfunktionen ℓ (IV) - Clustering

- Datensatz $\mathcal{D} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\}$ ohne Label
- Modell $\subset \mathbb{R}^n$
- k-Means (Intra-Cluster Varianz)

$$\mathrm{ICV}(k; \mathcal{D}) = \min_{C \subset \mathbb{R}^n, |C|=k} \sum_{\boldsymbol{x} \in \mathcal{D}} \min_{\boldsymbol{c} \in C} \|\boldsymbol{x} - \boldsymbol{c}\|_2^2$$

- Mixture Modelle (Expectation Maximization)
  - Example: Gaussian

$$\min_{\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{1 \le i \le k} \subset \mathbb{R}^n \times S^n_{++}} - \sum_{\boldsymbol{x} \in \mathcal{D}} \log \sum_{i=1}^{k} p_{\mathrm{Gauss}}(\boldsymbol{x} \mid i) p(i)$$