Seminar “Data Mining and Learning from Data”

Predictive Modeling

Thorsten Holz

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany

Predictive Modeling

- Grundbegriffe der Klassifikation und Regression
- Verschiedene Verfahren
  - Logistische Diskriminanzanalyse
  - Nächste-Nachbarn-Methode
  - Naive Bayes Modell
  - Andere Verfahren
- Evaluierung
  - Kreuzvalidierung / Bootstrap
  - Receiver-Operating-Characteristic (ROC) – Kurven
- Zusammenfassung

Hauptquelle:

- **D. Hand, H. Mannila and P. Smyth: „Principles of Data Mining“, MIT Press, Cambridge MA, 2001, (Kapitel 10 + 11).**

Sonstige Bücher:

- **R.O. Duda, P.E. Hart and D.J. Stock: „Pattern Classification“, Wiley, 2001.**
- **T. Hastie, R.J. Tibshirani and J. Friedman: „The Elements of Statistical Learning“, Springer, 2001.**
- **G.J. McLachlan: „Discriminant Analysis and Statistical Pattern Recognition“, Wiley, 1992.**

„We are drowning in information and starving for knowledge."
– Rutherford D. Roger

Immer mehr Daten werden gesammelt und gespeichert:

Einkaufsverhalten, Verbindungsdaten, astronomische Daten,...

```
Data-Mining is the analysis of (often large) observational data sets to find
unsuspected relationships and to summarize the data in novel ways that are
both understandable and useful to the data owner
```

- **Predictive Modeling macht Vorhersage über zukünftige Werte der Daten**
- **Modell: $y = f(\mathbf{x}, \Theta)$ mit Zielvariable y, Prädiktor $\mathbf{x} \in \mathbb{R}^D$, Parameter Θ**
- **Bei Klassifikation ist y kategorische Variable, bei Regression reellwertig**

Beispiel für Klassifikation:

Beurteilung Kreditwürdigkeit anhand verschiedener Variablen $\mathbf{x} = (x_1, \ldots, x_D)$,

bspw. „monatl. Einkommen“, „Höhe Gesamtschulden“,...

- **Zielvariable (engl.: response variable)** $y = f(\mathbf{x}, \Theta)$
  {**„Kredit gewährt“, „Kredit nicht gewährt“**}
- **Prädiktor (engl.: predictor variable)** $\mathbf{x} \in \mathbb{R}^D$
  („monatl. Einkommen“, „Familienstand“, „Gesamtschulden“,...)
- **Parameter des Modells** Θ
  Gewichtung der Prädiktor-Variablen

Lernen der Parameter Θ **durch Trainingsmenge** $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

Beispiele für Regression:

- **Mobilfunkkonzern möchte anhand von gewonnenen Verbindungsdaten Vorhersage über Kundenverhalten zu bestimmten Uhrzeiten treffen**
- **Vorhersage der Energieproduktion eines Windrads anhand von Wetterdaten u.ä.**

Lineare Modelle:

$$\hat{y} = a_0 + \sum_{d=1}^{D} a_d x_d$$

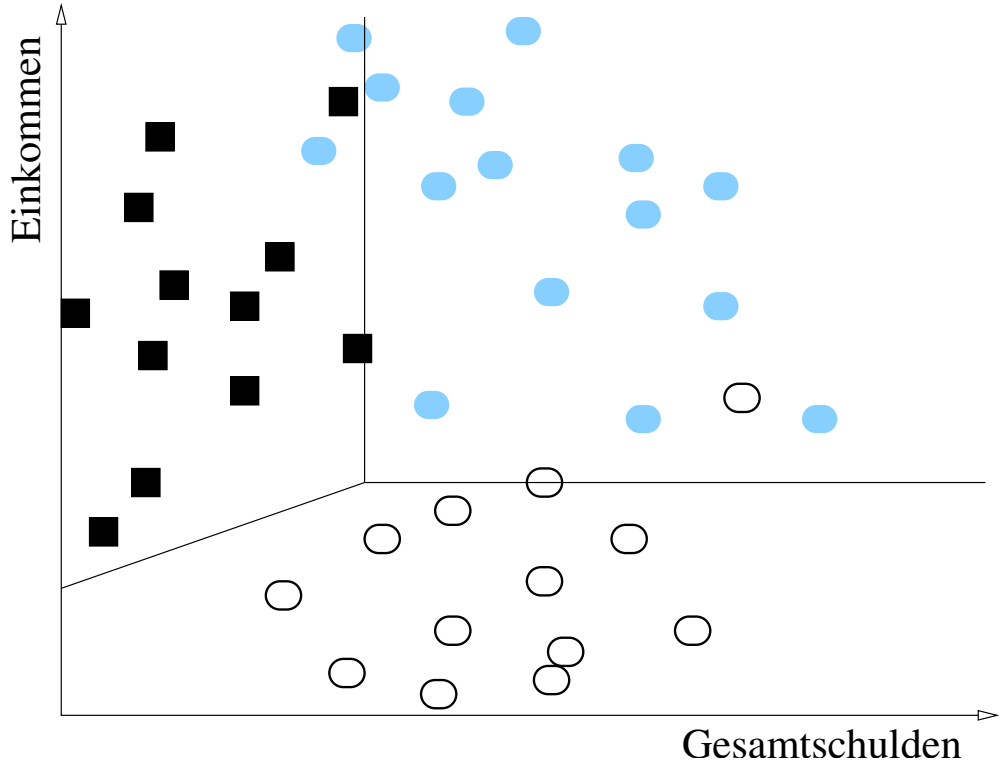
$\hat{y}$ ist durch Modell gewonnener Schätzer und a_d sind Gewichtungen

Schätzer ist fehlerbehaftet: $y = \hat{y} + \varepsilon$, **Residuum (engl.: residual)** ε

Minimierung des Residuums angestrebt, least-squares-Methode:

$$\sum_{n=1}^{N} \varepsilon(n)^2 = \sum_{n=1}^{N} \left(y(n) - \left(a_0 + \sum_{d=1}^{D} a_d(n)\ x_d(n) \right) \right)^2$$

Gegenüberstellung von „Höhe Gesamtschulden“ und „monatl. Einkommen“



- Drei verschiedene Risikotypen erkennbar
- Lineare Entscheidungsgrenze (engl.: decision boundary)
- Einige Werte falsch klassifiziert

- **Bias ist Maß für Genauigkeit; Abweichung des Erwartungswerts des Schätzers vom tatsächlichen Wert**

- **Varianz ist Maß für die Streuung; erwartete quadrierte Abweichung des Schätzers vom Erwartungswert des Schätzers**

Erwarteter Fehler einer Klassifikation $\hat{f}(\mathrm{x})$ bei Prädiktor x

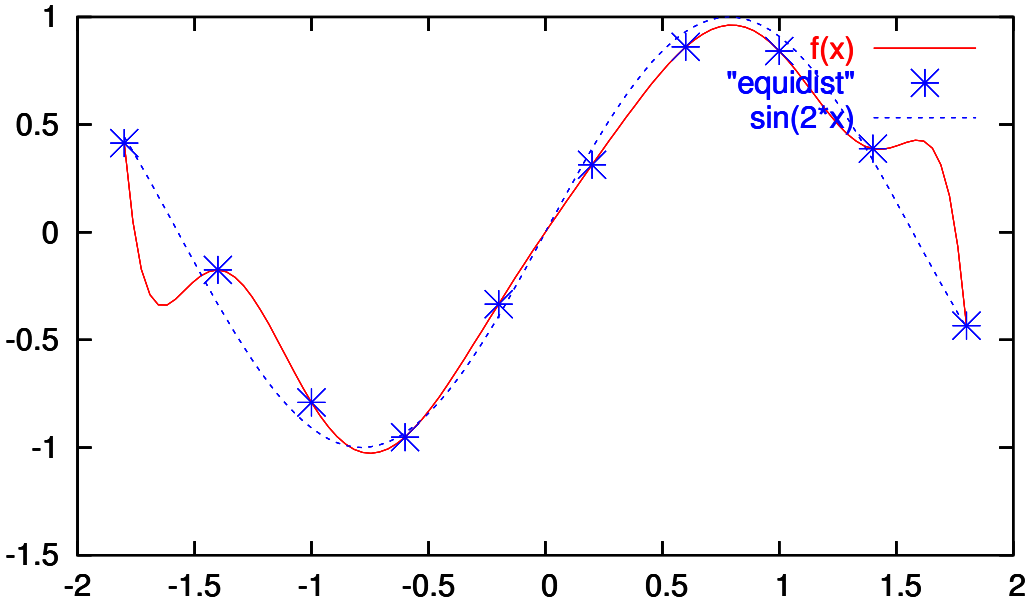
$$
\begin{aligned}
Err\{\mathrm{x}\} &= E\{(y - \hat{f}(\mathrm{x}))^2 | \mathrm{x}\} \\
&= \sigma_{\mathrm{x}}^2 + [E\{\hat{f}(\mathrm{x})\} - f(\mathrm{x})]^2 + E\{\hat{f}(\mathrm{x}) - E\{\hat{f}(\mathrm{x})\}\}^2 \\
&= \sigma_{\mathrm{x}}^2 + Bias^2\left(\hat{f}(\mathrm{x})\right) + Var\left(\hat{f}(\mathrm{x})\right) \\
&= \text{Minimaler Fehler} + \text{Bias}^2 + \text{Varianz}
\end{aligned}
$$

Beide Terme können nicht gleichzeitig minimiert werden!

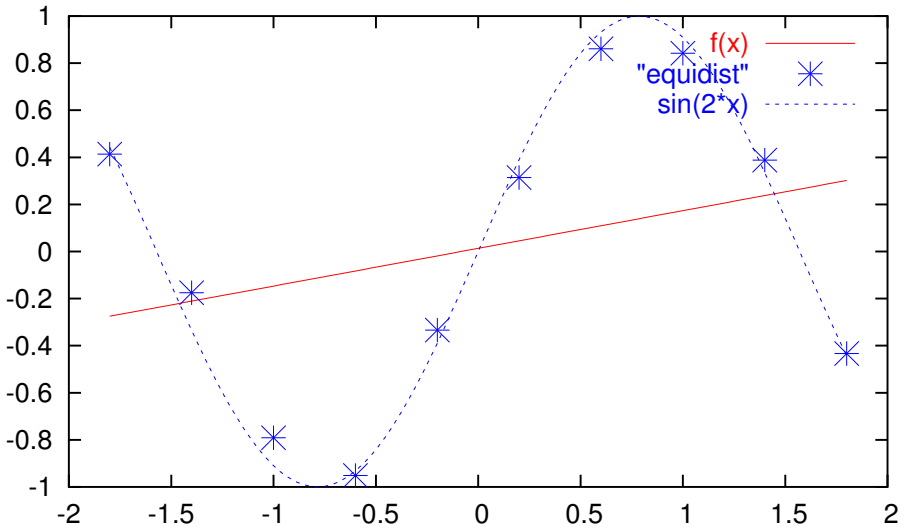
- **Funktion $\sin(2x)$ im Bereich [-2,2]**
- **Trainingsdaten mittels Funktion $\sin(2\mathrm{x}) + \varepsilon$ mit $\varepsilon \sim \mathcal{N}(0, \frac{1}{9})$**

Least-Squares-Methode mit 10 Freiheitsgraden ($f_{10}(\mathrm{x}) = \sum_{i=0}^{9} a_i x^i$)

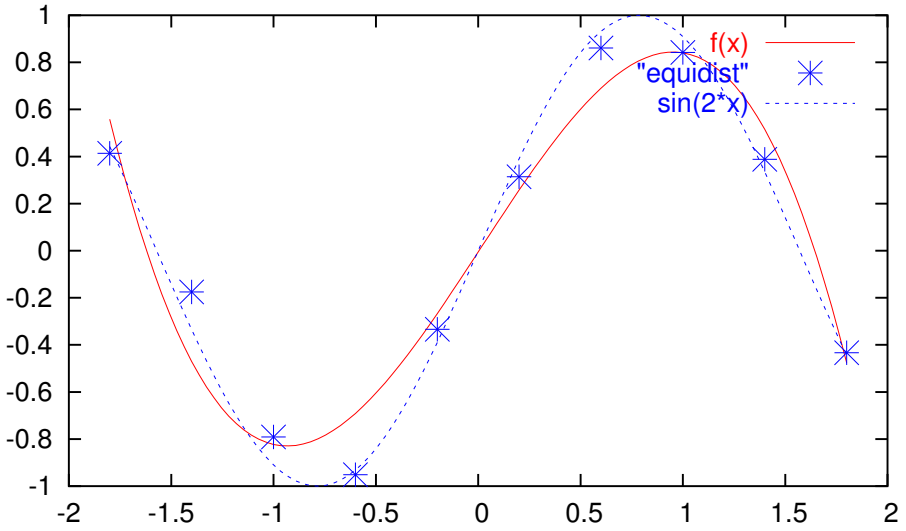
„overfitting“, hohe Varianz, aber niedriger Bias



Least-Squares Methode mit 2 Freiheitsgraden,
($f_2(\mathbf{x}) = a_0 + a_1 x$)
„underfitting“, niedrige Varianz, hoher Bias

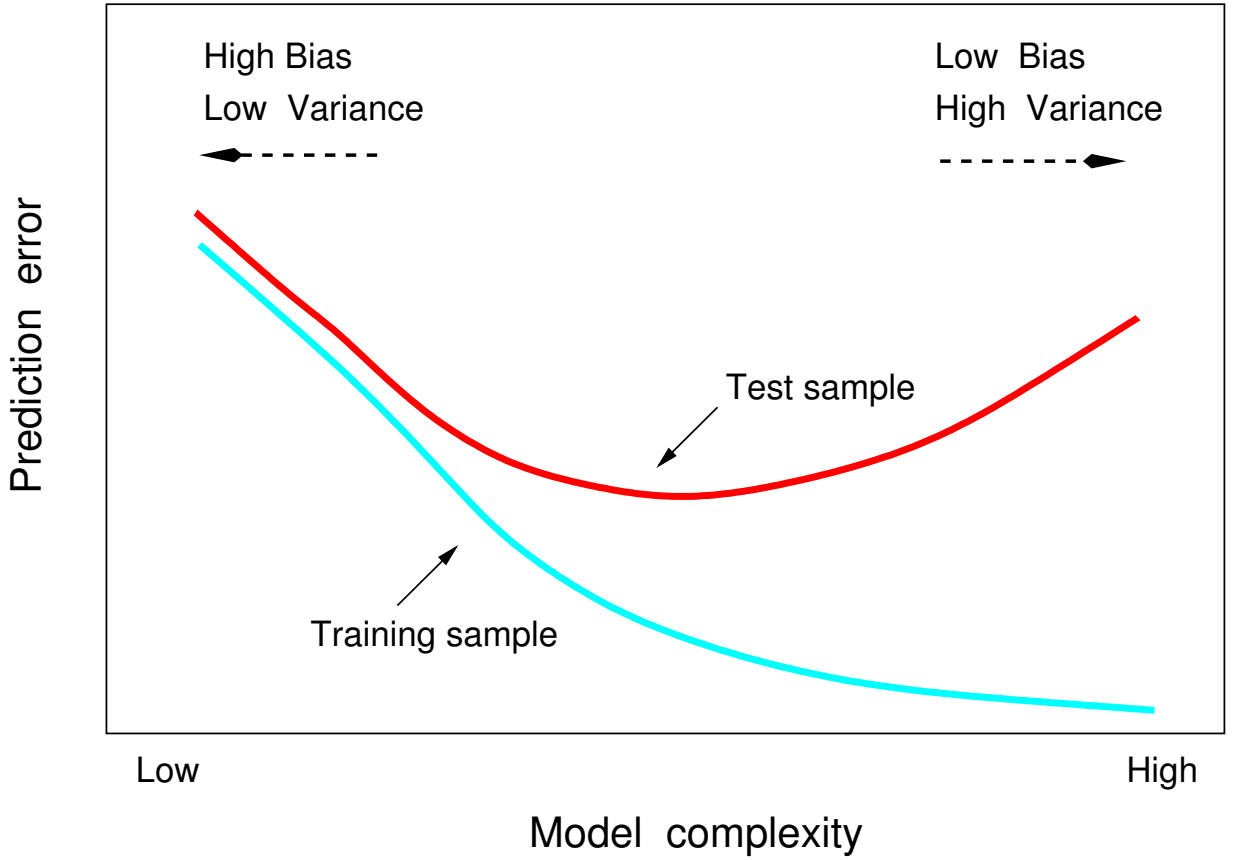


Least-Squares Methode mit 4 Freiheitsgraden,
($f_4(\mathbf{x}) = \sum_{i=0}^{3} a_i x^i$)
gute Annäherung an die Daten



Zusammenhang zwischen Modellkomplexität und Vorhersagefehler

- **Trade-Off zwischen Bias und Varianz**
- **Oft adaptiver Ansatz nötig**



Verallgemeinerte lineare Modelle: Zielvariable y ist indirekt über Link-Funktion $g(y)$ von Linearkombination der x_d abhängig:

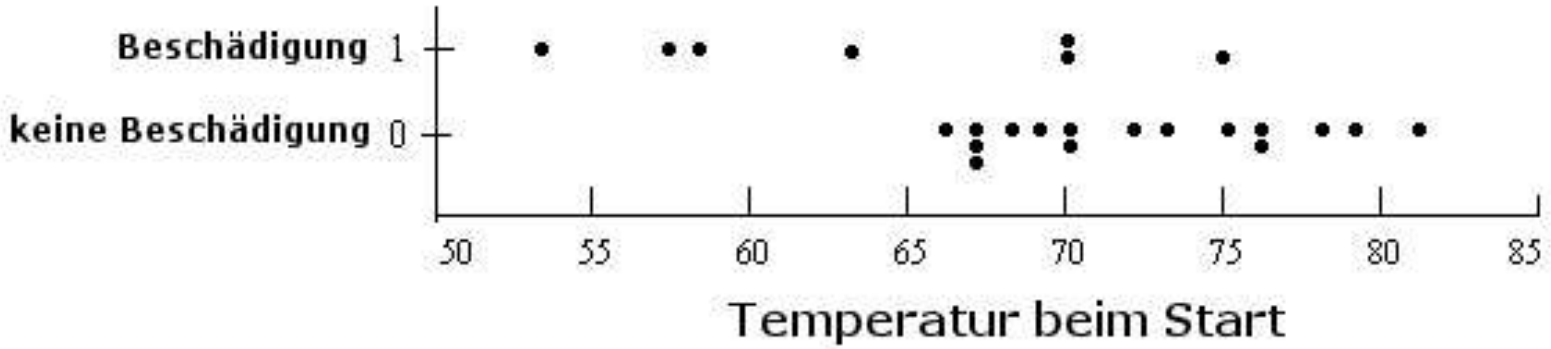
$$g(y) = a_0 + \sum_{d=1}^{D} a_d x_d$$

- Logistische Diskriminanzanalyse nutzt als Link-Funktion Logarithmus (log-lineares Modell)
- Einfachster Fall: Vorhersage einer binären Zielvariablen
  - Wahrscheinlichkeit für Klasse $c = 1$ bei Prädiktor x
    $$p(c = 1|\text{x}) = \frac{1}{1+\exp(\alpha^T \text{x'})}$$
  - Gegenereignis:
    $$p(c = 0|\text{x}) = 1 - p(c = 1|\text{x}) = \frac{\exp(\alpha^T \text{x'})}{1+\exp(\alpha^T \text{x'})}$$
  - odds-ratio: Verhältnis der relativen Häufigkeiten möglicher Ereignisse

$$\log\left(\frac{p(c = 0|\text{x})}{p(c = 1|\text{x})}\right) = a_0 + \sum_{d=1}^{D} a_d x_d$$

28. Januar 1986: Raumfähre Challenger explodiert beim Start

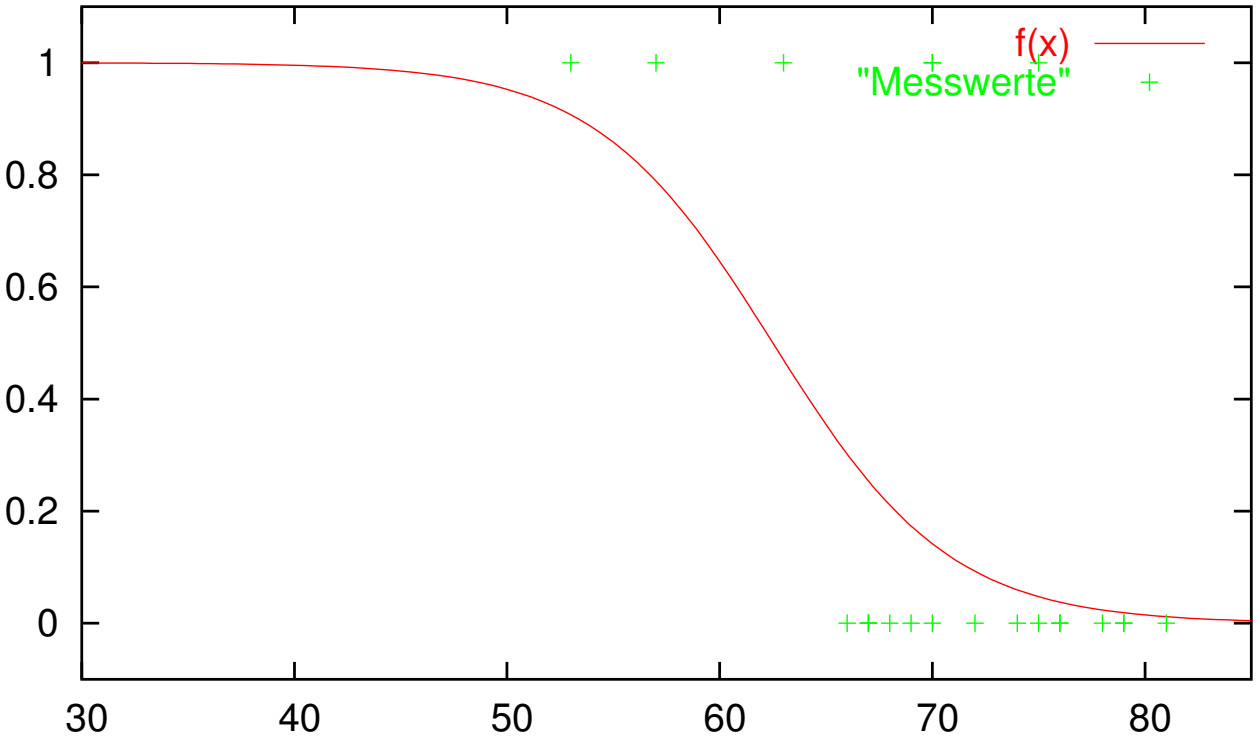
- **Brüchige Gummiringe an Antriebsrakete unmittelbare Ursache**
- **Vorhergesagte Temperatur für 28. Januar: 31° Fahrenheit**



Anwendung der logistischen Diskriminanzanalyse zur Modellbildung

$$f(\mathbf{x}, \Theta) = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + \exp(\alpha_0 + \alpha_1 x)} = \frac{\exp(15 - 0.25x)}{1 + \exp(15 - 0.25x)}$$

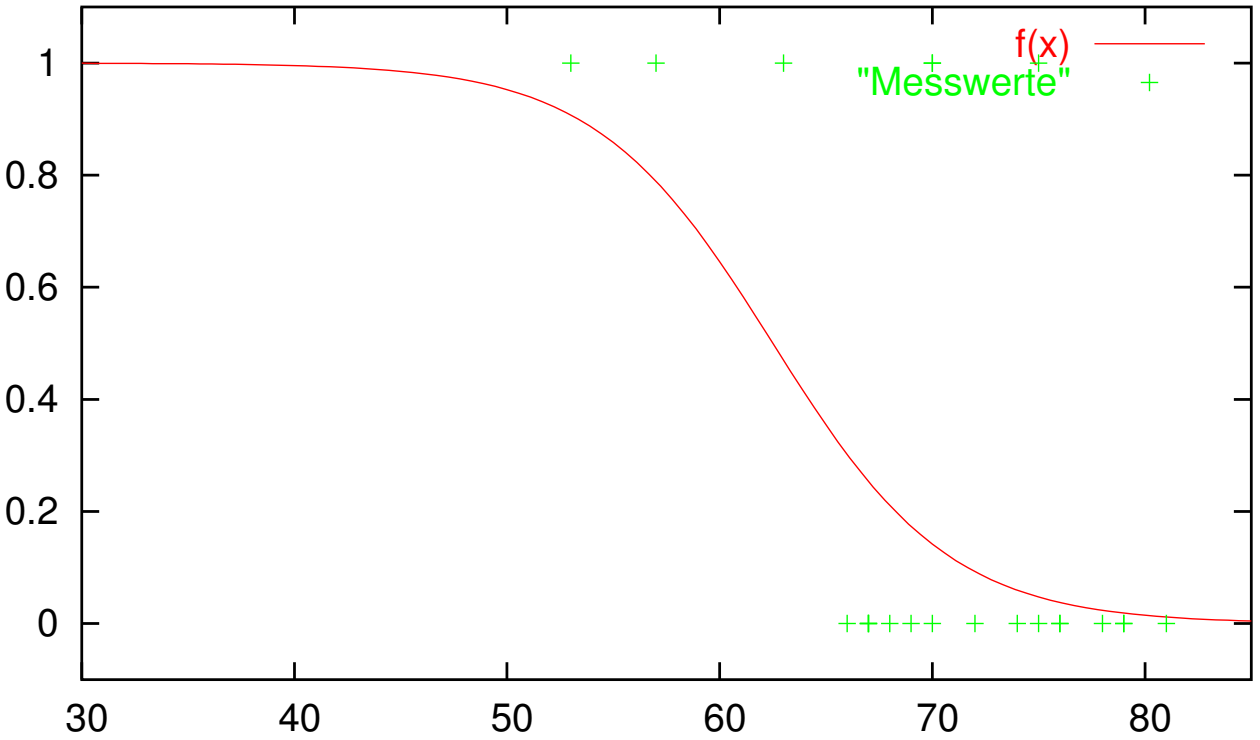
Logistische Diskriminanzanalyse: Analyse des Challenger-Unglücks



Anwendung der logistischen Diskriminanzanalyse zur Modellbildung

$$f(\mathbf{x}, \Theta) = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + \exp(\alpha_0 + \alpha_1 x)} = \frac{\exp(15 - 0.25x)}{1 + \exp(15 - 0.25x)}$$

Logistische Diskriminanzanalyse: Analyse des Challenger-Unglücks



Wahrscheinlichkeit für Beschädigung der Gummiringe anhand dieses Modells lag nahe bei 1

$\Rightarrow$ Start hätte nicht stattfinden dürfen

Objekt mit Prädiktor $\mathbf{x}'$ soll einer von $c = 1, \ldots, C$ Klassen zugeordnet werden

1. Festlegung eines Distanzmaßes
   - Minkowski-Distanz:

$$\mathbf{dist_p}(\mathbf{x}', \mathbf{x}) = \left(\sum_{d=1}^{D} |x'_d - x_d|^p \right)^{\frac{1}{p}}$$

     Wichtige Spezialfälle:

     $p = 1$ Manhattan-Distanz und $p = 2$ euklidische Distanz

   - Mahalanobis-Distanz:

$$\mathbf{dist_{mah}}(\mathbf{x}', \mathbf{x}) = (\mathbf{x}' - \mathbf{x})^T \Sigma^{-1} (\mathbf{x}' - \mathbf{x})$$

     Verallgemeinerung der euklidischen Distanz unter Berücksichtigung der Korrelationen der x, Parameter ist Kovarianzmatrix Σ

2. Überprüfen der k nächsten Nachbarn von $\mathbf{x}'$

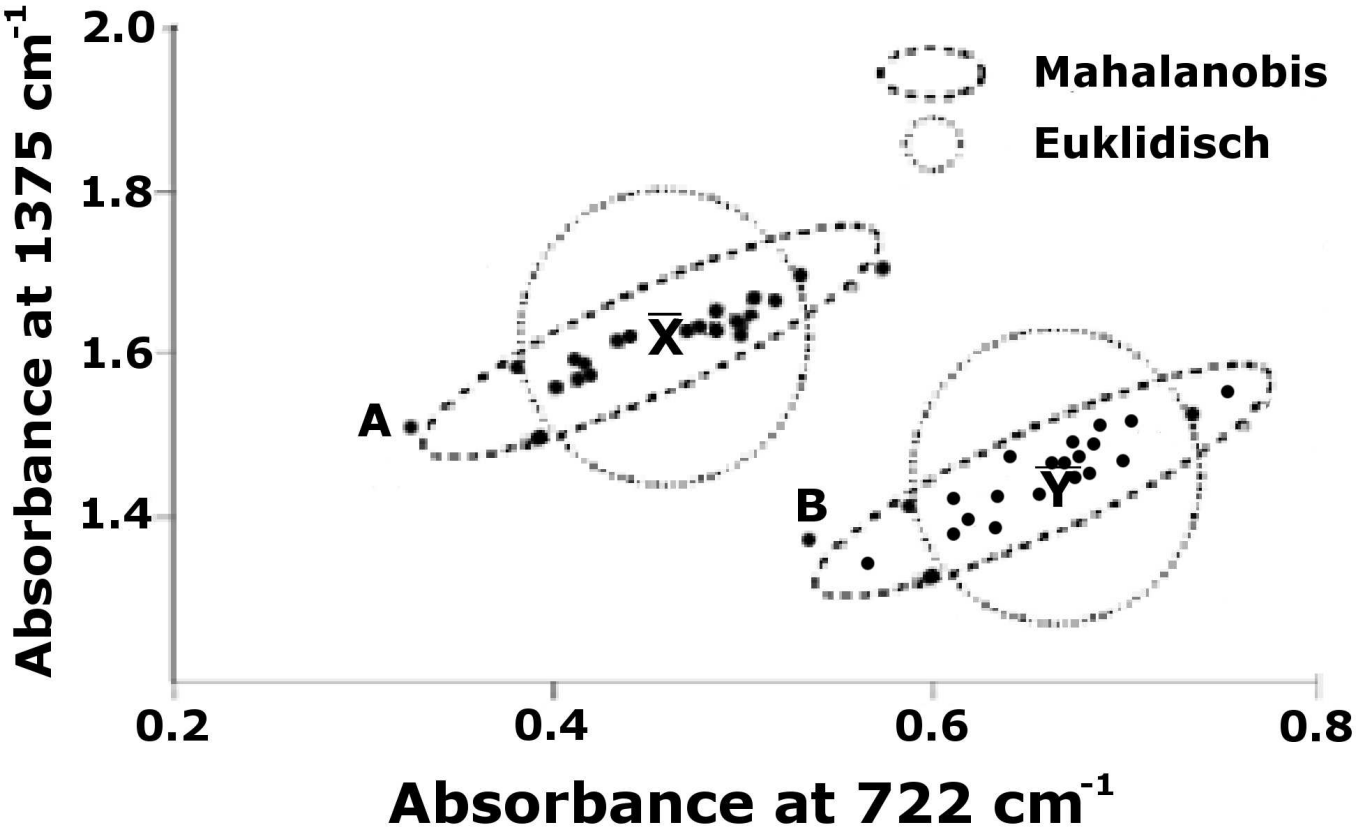
3. Zuordnung von $\mathbf{x}'$ zu genau der Klasse c^*, die bei den k Nachbarn am häufigsten auftritt

Doch wie groß soll k gewählt werden?

- $k = 1$ liefert unstabilen Klassifikator mit hoher Varianz
- Erhöhung von k bewirkt Stabilisierung:
  Varianz nimmt aber, allerdings kann Bias erhöht werden
- Wahl hängt stark von Daten ab $\Rightarrow$ adaptiver Ansatz

Messung der Absorption bei verschiedenen Stoffen bei verschiedenen Wellenlängen

- **Verwendung zweier Distanzmaße**
- **Mittelwert der gemessenen Punkte ist $\overline{x}$ bzw. $\overline{y}$**



Vor- und Nachteile der Nächste-Nachbarn-Methode

+ Einfache Implementierung, intuitive Methode

+ Einfaches Erkennen von Ausreißern:
  Distanz zu Nachbarn liegt über definiertem Schwellenwert

+ Bei fehlenden Prädiktor-Variablen Durchführung im Unterraum

– Speicherverbrauch hoch, da alle Trainingsdaten gespeichert werden müssen

– Laufzeit bei komplizierten Nachbarschaften oder großer Anzahl Trainingsdaten schlecht

– Dünn besetzte Nachbarschaften bei großer Dimension des Prädiktors

Satz von Bayes: $p(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)p(c_i)}{p(\mathbf{x})}$

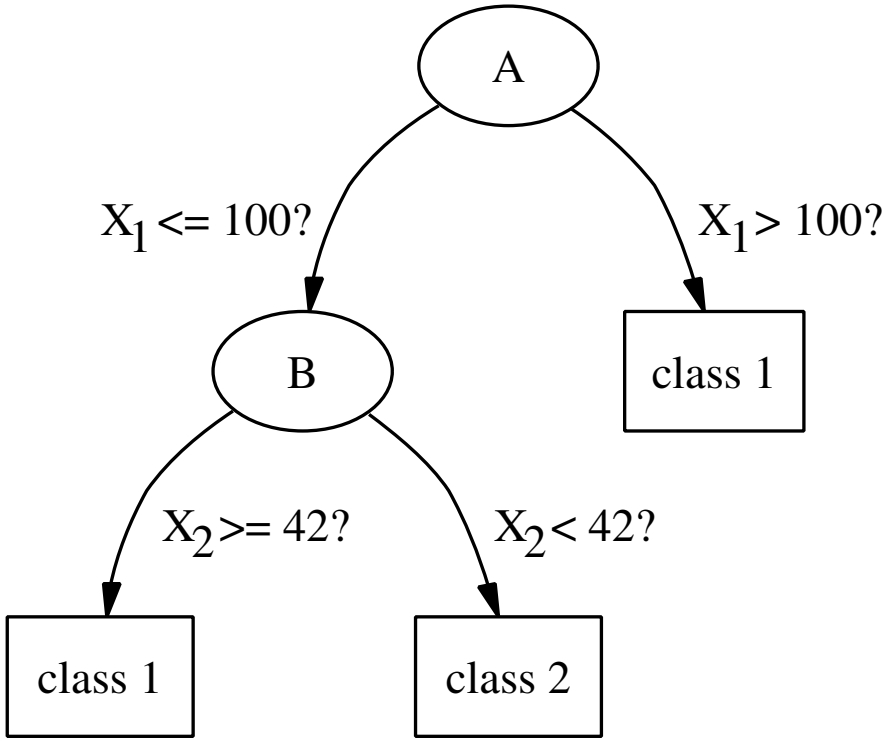
- **Naives Bayes Modell, da die stochastische Unabhängigkeit der Prädiktor-Variablen angenommen wird**

$$p(\mathbf{x}|c_i) = p(x_1, \ldots, x_D|c_i) \cong \prod_{d=1}^{D} p(x_d|c_i), \qquad i = 1, \ldots, C$$

- **Annahme i.A. falsch, da innerhalb Prädiktor-Variablen Abhängigkeit auftreten kann**
- **In vielen praktische Fällen dennoch möglich, starke Vereinfachung des Verfahrens**
- **Benutzung der höchsten auftretenden Wahrscheinlichkeit zur Klassifikation (Bayes'sche Entscheidungsregel)**

- **Spezielle Form der nichtlinearen Diskriminanzanalyse**
- **Benutzt Bäume zur Klassifikation:**

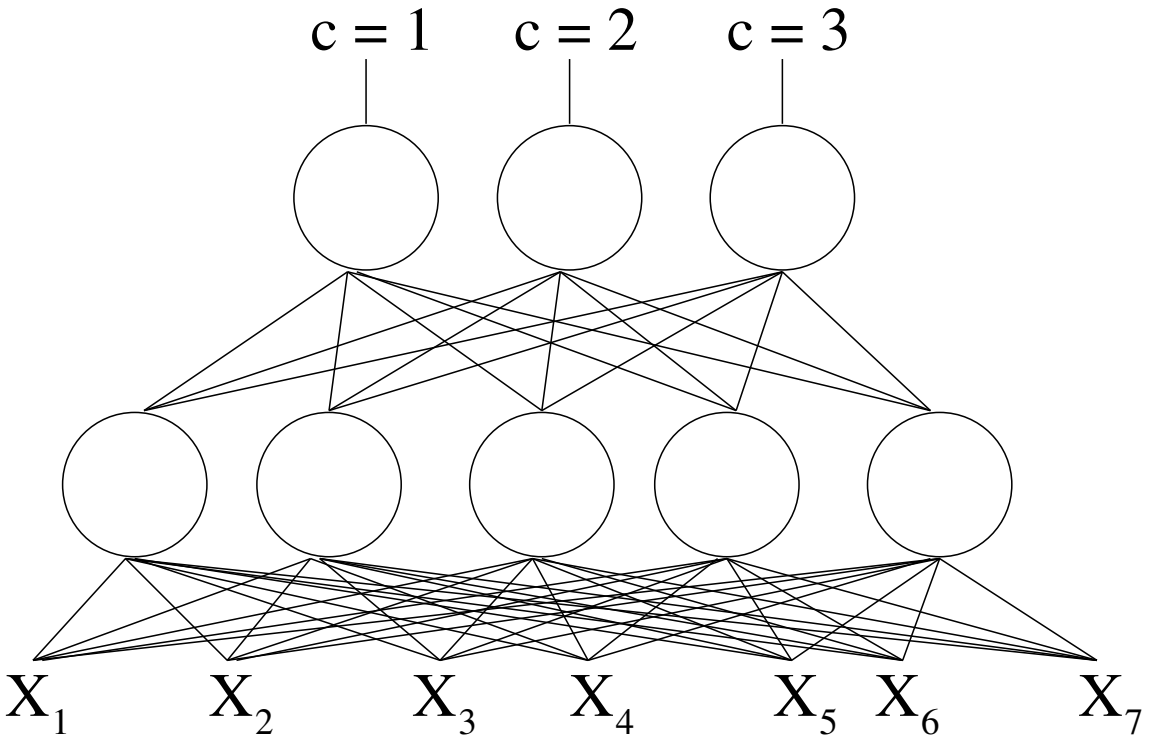
  Blätter repräsentieren Klassen, Kanten bilden Entscheidungen
- **Konstruktion per pruning-Methode**



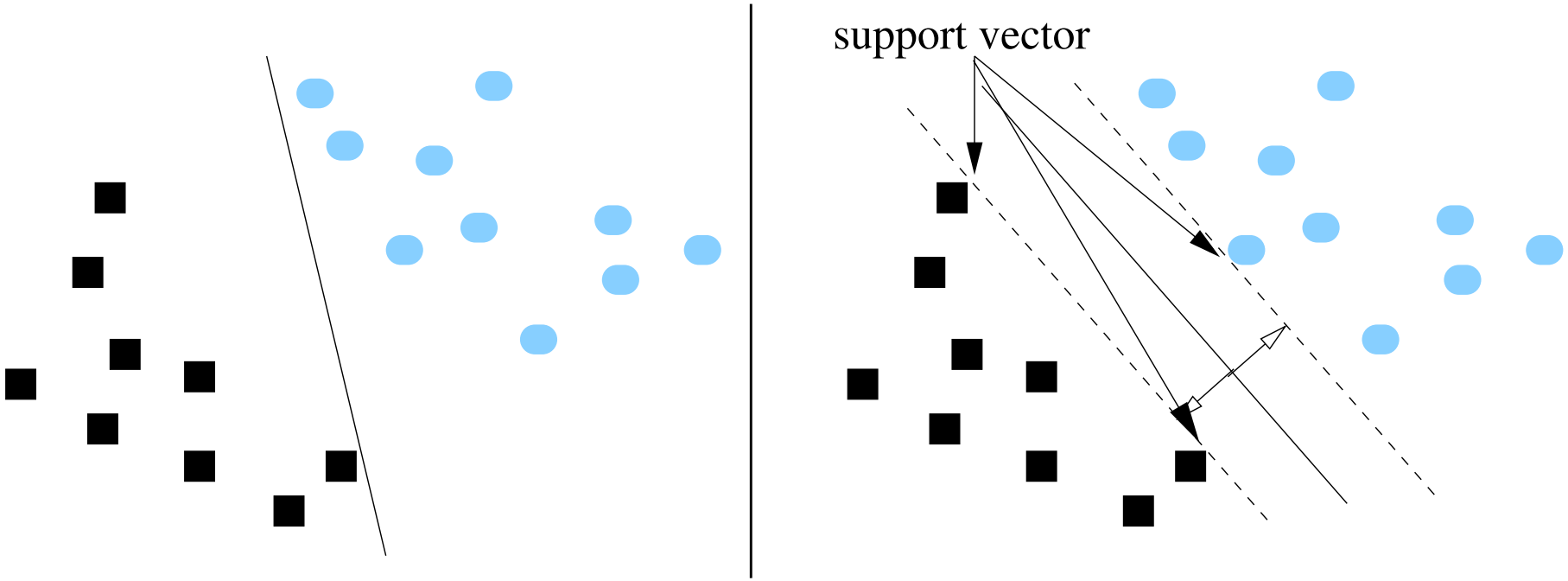
- **Idee ähnlich der Neuronen im Gehirn:**

  Ab bestimmtem Schwellenwert wird Signal an Ausgang weitergeleitet

- **Lernen der Kantengewichtungen durch Training**



- Nichtlineare Transformation in höherdimensionalen Raum
- Suche nach Hyperebenen mit möglichst guter Trennung der Trainingsdaten
- Support-Vektoren liegen am nächsten an optimal trennender Hyperebene



Wie kann man die Leistungsfähigkeit eines Modells abschätzen?

- **Statistische Kennzahlen, beispielsweise „Anzahl falsch klassifizierter Testdaten“ oder „Fehlerrate“**
- **Berücksichtigung der Kosten für falsch klassifizierte Objekte**
- **Minimum description length:**
  - **Modell M mit kürzester Beschreibungslänge (basierend auf Entropie nach Shannon) wird als Bestes angenommen**
  - **length$(M) = -\log p(y|\Theta, M, \mathbf{x}) - \log p(\Theta|M)$**
  - **1. Term: durchschnittliche Länge der Kodierung der Abweichung von $\hat{y}$ zu y**
  - **2. Term: durchschnittliche Länge der Kodierung von Θ**

Einfachste und am häufigsten benutze Methode zur Schätzung des Fehlers:

Kreuzvalidierung (engl.: cross-validation)

- **Zerlegung der Gesamtmenge in N gleich große Mengen**
- **$N-1$ Mengen zum Training, eine Testmenge**

$\Rightarrow$ **N-fache Wiederholung, Kombinierung der Ergebnisse**

Wie groß sollte N gewählt werden?

- **leave-one-out-Methode: Aufteilung mit $N = |$Gesamtmenge$|$**
  Größtmögliche Trainingsmenge, hoher Rechenaufwand
- **Bei praktischen Anwendungen oft $N = 5$ und $N = 10$**

Allgemeines Verfahren zur Berechnung von statistischen Schwankungen

- N zufällige Stichproben (bootstrap-samples) mit Zurücklegen aus Trainingsmenge ziehen
  ($\longrightarrow$ einige Elemente können mehrfach vorkommen)
- Generierung weiterer solcher bootstrap-Datensätze (beispielsweise 100 Stück)
- Training und Test mittels dieser bootstrap-Datensätze
  ($\longrightarrow$ Problem des „training on the testing data“)

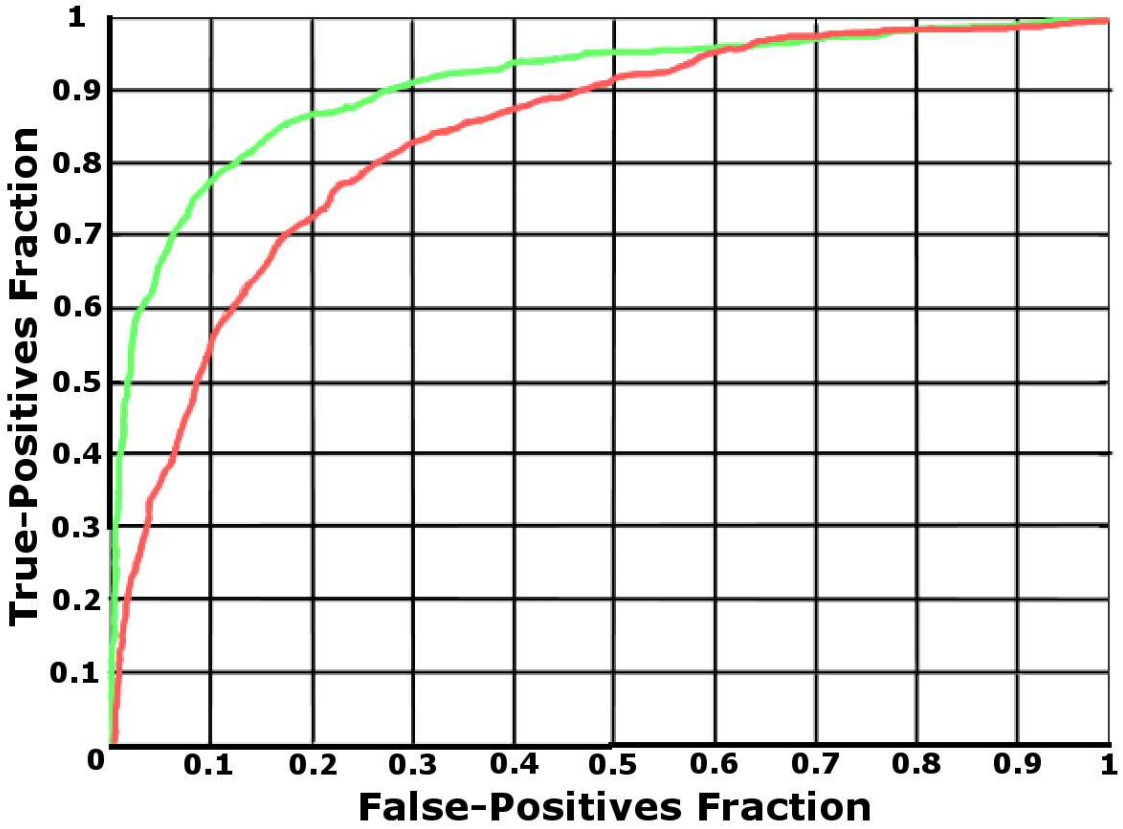
Leave-one-out-bootstrap:

- Vermeidung der zu positiven Fehlerabschätzung
- Test nur mit denjenigen Datensätzen, bei denen ein bestimmtes $(\mathbf{x}_n, y_n)$ nicht im bootstrap-Datensatz enthalten ist:

  $$(1 - \frac{1}{N})^N \approx \mathrm{e}^{-1} \approx 0,368$$

Receiver-Operating-Characteristic (ROC) – Kurven

- Visualisierung der Sensitivität und Spezifität von Modellen im Zwei-Klassen-Fall
- Gegenüberstellung des False – und True - Positives Anteils



Quelle: http://vision.ai.uiuc.edu/myhang/face-detection-survey.html

- ✔ **Predictive Modeling:**

  Vorhersage zukünftiger Werte von Daten anhand gegebener Datenmenge

- ✔ **Vielzahl von Verfahren möglich**
  - **Logistische Diskriminanzanalyse**
  - **Nächste-Nachbarn-Methode**
  - **Naive Bayes Modell**
  - **...**

- ✔ **Auswahl des Verfahrens abhängig von Art der Daten**

- ✔ **Aussagen über Leistungsfähigkeit der Verfahren beispielsweise mittels**
  - **Kreuzvalidierung / Bootstrap**
  - **Receiver-Operating-Characteristic – Kurven**

- ✔ **Bias-Varianz-Problem immer beachten**

Vielen Dank für ihre Aufmerksamkeit!

$$\sum = \begin{pmatrix} Var(x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & Cov(x_1, x_m) \\ Cov(x_2, x_1) & Var(x_2) & \cdots & Cov(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_m, x_1) & Cov(x_m, x_2) & \cdots & Var(x_m) \end{pmatrix}$$

$Cov(x_i, x_j)$ bezeichnet Kovarianz (engl.: covariance) von x_i und x_j und ist ein Maß für Abhängigkeit von x_i und x_j:

$$Cov(x_i, x_j) = E\{(x_i - E\{x_i\}) \cdot (x_j - E\{x_j\})\}$$

$Var(x_i)$ bezeichnet Varianz (engl.: variance) der i-ten Komponente. Maß für Streuung und gibt erwartete quadrierte Abweichung vom Erwartungswert an:

$$Var(x_i) = Cov(x_i, x_i) = \sigma_i^2 = E\{(x_i - E\{x_i\})^2\}.$$

Einzelner Wert y mit $y \sim \mathcal{N}(\Theta, \sigma^2)$, Modellparameter $\Theta \sim \mathrm{N}(0,1)$ und ohne Prädiktor:

$$
\begin{aligned}
\mathbf{length(M)} &= -\log \mathrm{p}(y|\Theta, \mathrm{M}) - \log \mathrm{p}(\Theta|\mathrm{M}) \\
&= -\log\left(\frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(y-\Theta)^2}{2\sigma^2}\right)\right) - \log\left(\frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{\Theta^2}{2}\right)\right) \\
&= -\log\left(\frac{1}{\sqrt{2\pi}}\right) + \log(\sigma) - \log\left(\exp\left(-\frac{(\mathrm{y}-\Theta)^2}{2\sigma^2}\right)\right) \\
&\quad -\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\left(\exp\left(-\frac{\Theta^2}{2}\right)\right) \\
&= \mathrm{const} + \log(\sigma) + \frac{(y-\Theta)^2}{2\sigma^2} + \frac{\Theta^2}{2}
\end{aligned}
$$