

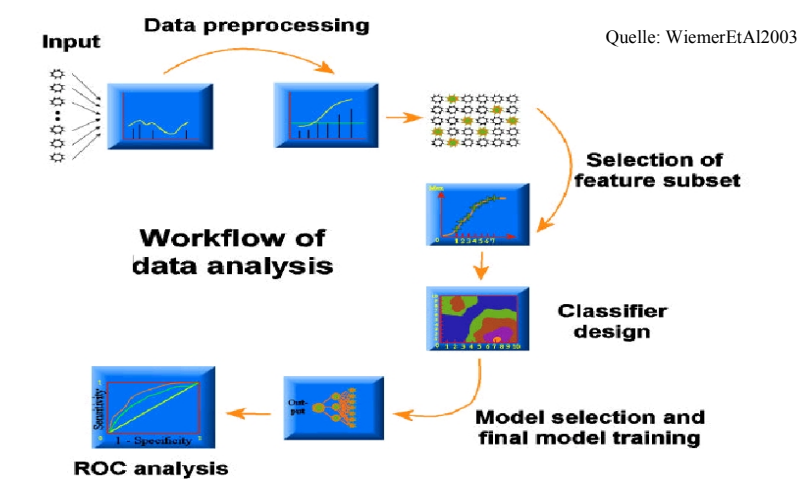
DKFZ Heidelberg

# Maschinelles Lernen, Klassifikation und diagnostische Tests (3)

Falk Schubert  
Intelligente Bioinformatiksysteme  
DKFZ Heidelberg

f.schubert@dkfz.de

## Data-Mining-Prozess



# Diagnosetests

---

## Evaluation von Klassifikatoren

# Vierfeldertafel (confusion matrix)

---

Aufstellung aller möglichen Testergebnisse:

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

# Sensitivität / Spezifität

---

$$\text{Sensitivität} = P(\text{Test positiv} \mid \text{krank}) = A/(A+C)$$

$$\text{Spezifität} = P(\text{Test negativ} \mid \text{gesund}) = D/(B+D)$$

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

# Sensitivität

---

$$\text{Sensitivität} = P(\text{Test positiv} \mid \text{krank}) = A/(A+C)$$

Wahrscheinlichkeit, dass Test bei kranken Personen positiv ausfällt.

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

# Spezifität

$$\text{Spezifität} = P(\text{Test negativ} \mid \text{gesund}) = D/(B+D)$$

Wahrscheinlichkeit, dass Test bei gesunden Personen negativ ausfällt.

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

# Vorhersagewerte

$$\begin{aligned} \text{Positiver Vorhersagewert (positive predictive value, ppv)} \\ = P(\text{krank} \mid \text{Test positiv}) = A/(A+B) \end{aligned}$$

$$\begin{aligned} \text{Negativer Vorhersagewert (negative predictive value, npv)} \\ = P(\text{gesund} \mid \text{Test negativ}) = D/(C+D) \end{aligned}$$

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

## Positiver Vorhersagewert

---

Positiver Vorhersagewert (positive predictive value, ppv)  
=  $P(\text{krank} | \text{Test positiv}) = A/(A+B)$

Wahrscheinlichkeit, dass eine Person mit einem positiven Test tatsächlich erkrankt ist.

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

## Negativer Vorhersagewert

---

Negativer Vorhersagewert (negative predictive value, npv)  
=  $P(\text{gesund} | \text{Test negativ}) = D/(C+D)$

Wahrscheinlichkeit, dass eine Person mit einem positiven Test tatsächlich erkrankt ist.

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

## Wiederholung Satz von Bayes

---

Grundlegend für die Bestimmung bedingter  
Wahrscheinlichkeiten

Version für zwei Ereignisse:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A | B)P(B) + P(A | \bar{B})P(\bar{B})}$$

## Prävalenz / Inzidenz

---

- Prävalenz = P (krank)
- Anteil der Kranken an Population
  
- Inzidenz
- Anteil neu Erkrankter an einer Population innerhalb eines Zeitraumes

## Berechnung PPV mit Bayes

---

Positiver Vorhersagewert (PPV)

$$= P(\text{Krank}|\text{Test+})$$

$$= \frac{P(\text{Test+}|\text{Krank}) * P(\text{Krank})}{P(\text{Test+}|\text{Krank}) * P(\text{Krank}) + P(\text{Test+}|\text{Gesund}) * P(\text{Gesund})}$$

$$= \frac{\text{Sensitivität} * \text{Prävalenz}}{\text{Sensitivität} * \text{Prävalenz} + (1 - \text{Spezifität}) * (1 - \text{Prävalenz})}$$

## Beispiel HIV-Test (1)

---

- ELISA-Antikörper-Test  
(enzyme-linked immunosorbent assay)
- HIV-Antikörper-Test positiv
- Spezifität: 98.3 %
- Sensitivität: 95.9 %  
(Lancet. 1992 Dec 19-26;340(8834-8835):1496-9)
- Wahrscheinlichkeit dafür, dass Patient tatsächlich HIV infiziert ist?

## Beispiel HIV-Test (2)

---

- Prävalenz 0.05% (Gesamtbevölkerung)  
ppV = 2.7%
- Prävalenz 1% (men having sex with men)  
ppV = 36%
- Prävalenz 14% (Kenia)  
ppV = 90%

Hinweis: Prävalenzen sind geschätzt.

## Beispiel HIV-Test (3)

---

- Praktische Konsequenz:  
  
Bestätigung mit Western-Blot,  
ggf. zweite Blutentnahme
- Spezifität und Sensitivität des Elisa-Tests  
mittlerweile 99.5% (-> Übungsaufgabe)

## Bewertung von Klassifikatoren

---

### Beurteilung des Klassifikators

---

- Wie gut kann der Klassifikator zukünftige Daten klassifizieren?
- Vergleich von Klassifikatoren
- Rückschluss auf die Separierbarkeit von Klassen
  - 2 Tumore sind aufgrund ihrer genomischen Profile unterscheidbar
- Optimierung unter Berücksichtigung der Kosten von (Fehl)-Entscheidungen

## Trainings- und Generalisierungsfehler

---

Zielkriterium, Generalisierungsfehler

$$Err = E[L(Y, \hat{f}(X))]$$

Trainingsfehler, Reklassifikationsfehler

$$err = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$$

## Verlustfunktionen

---

0/1 Fehler

$$L(y_i, \hat{f}(x_i)) = \begin{cases} 0, & y_i = \hat{f}(x_i) \\ 1, & y_i \neq \hat{f}(x_i) \end{cases}$$

Quadratischer Fehler

$$L(y_i, \hat{f}(x_i)) = (y_i - \hat{f}(x_i))^2$$

Absoluter Fehler

$$L(y_i, \hat{f}(x_i)) = |y_i - \hat{f}(x_i)|$$

# Reklassifikation

---

- Trainiere einen Klassifikator mit den Trainingsdaten
- Evaluire den Klassifikator auf diesen Trainingsdaten
  
- Optimistische Schätzung des Fehlers
- Eine objektive Evaluierung muss stets auf einem unabhängigen Datensatz erfolgen!

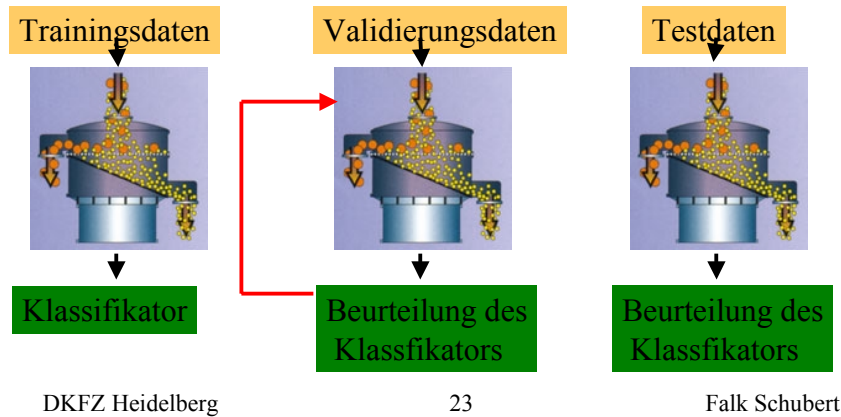
# Design

---

- Daten
  - Trainingsdaten
  - Testdaten
- Daten
  - Trainingsdaten
  - Testdaten
  - Validierungsdaten
    - Wahl des Klassifikators
    - Wahl der Parameter des Klassifikators

## Testdaten

- Datenmenge, die bei der Entwicklung des Klassifikators nicht berücksichtigt wurden



## Aufteilung Trainings-/Testdaten

- Zufällig
- Stratifiziert
- Mehr Trainingsdaten
  - Besserer Klassifikator, weniger Bias
- Mehr Testdaten
  - Genauere Fehlerabschätzung, weniger Varianz
- Bei größeren Datensätzen
  - 2/3 Trainingsdaten, 1/3 Testdaten
  - 50% Training, 25% Validierung, 25% Test

## Typische Designfehler

---

- Merkmalsauswahl wird optimiert für den Testdatensatz
  - Korrekt: Merkmalsauswahl muss auf Trainingsdatensatz optimiert werden.
- Wahl des Klassifikators wird abhängig gemacht von den Testergebnissen auf dem Testdatensatz
  - Korrekt: Wahl des Klassifikators a priori oder durch den Validierungsdatensatz.

## Typische Designfehler (2)

---

- Vorverarbeitung der Merkmale wird optimiert für den Testdatensatz
  - Korrekt: Vorverarbeitung muss optimiert werden für den Validierungsdatensatz.
- Test- und Trainingsdatensatz wurden nicht zufällig bestimmt.
  - Korrekt: Der Testdatensatz muss zufällig aus allen Daten gezogen werden.

## Kleine Datensätze

---

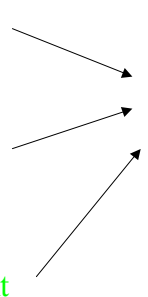
- Typisches Problem:
  - Zur Verfügung stehende Datensatz (z.B.  $n < 300$ ) erlaubt keine Unterteilung in Test-, Trainings- und Validierungsdatsatz.
- Wiederholte Training- und Test mit verschiedenen Stichproben
  - Bootstrap (Ziehen mit Zurücklegen)
  - Kreuzvalidierung (Ziehen ohne Zurücklegen)
  - LOO-Kreuzvalidierung

## LOO-Kreuzvalidierung

---

- Leave-one-out (LOO) cross validation
- Für alle  $n$  Datensätze
  - Verwende den aktuellen Datensatz als Testdatensatz
  - Verwende die restlichen  $(n-1)$  Datensätze für das Training
- Mittlere Ergebnisse über alle  $n$  Testdatensätze

## Beispiel LOO-CV

- Datensätze 1,2,3
  - Durchlaufe alle Datensätze
    - Datensatz 1
      - Training 2,3
      - Test 1 **Korrekt**
    - Datensatz 2
      - Training 1,3
      - Test 2 **Falsch**
    - Datensatz 3
      - Training 1,2
      - Test 3 **Korrekt**
- Mittlerer Fehler: 0.33
- 

## K-fache Kreuzvalidierung

- Aufteilung der Daten in K Partitionen gleicher Größe  $\kappa : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$
- Für alle Partitionen
  - Nutze eine Partition für den Test
  - Nutze die restlichen Partitionen für das Training
- 5 oder 10 Partitionen meist typisch

$$KV = \frac{1}{N} \sum_{i=1}^N L(Y_i, \hat{f}^{-\kappa(i)}(X_i)) \quad \hat{f}^{-k}(x) \text{ Klassifikator,}$$

trainiert ohne die Partition k

## Beispiel 3-fache CV

---

- Datensätze 1,2,3,4,5,6,7,8,9
- Durchlaufe alle drei Partitionen    Mittlerer Fehler: 20%
  - Partition1
    - Training: Partitionen 2,3 (Datensätze 4,5,6,7,8,9 )
    - Test: Partition 1 (Datensätze 1,2,3)    **Fehler =20%**
  - Partition 2
    - Training: Partitionen 1,3 (Datensätze 1,2,3,7,8,9 )
    - Test: Partition 2 (Datensätze 4,5,6)    **Fehler =15%**
  - Partition 3
    - Training: Partitionen 1,2 (Datensätze 1,2,3,4,5,6)
    - Test: Partition 3 (Datensätze 7,8,9)    **Fehler =25%**

## LOO-CV

---

- n-fache Kreuzvalidierung (n Anzahl Datensätze)
- Deterministisches Ergebnis, keine zufällige Wahl der Partitionen

## Fehler bei LOO-CV

---

- Konstruiertes Beispiel mit Fehlerüberschätzung
- Zwei Klassen, zufällige Klasseneinteilung, beide Klassen gleich mächtig
- Klassifikationsfehler: 50%
- Fehlerschätzung mit LOO: 100%

## 0.632-Bootstrap

---

- Ziehe eine Stichprobe von  $n$  Datensätzen mit Zurücklegen aus  $n$  Datensätzen
- Stichprobe enthält 63.2 % der ursprünglichen Datensätze
$$p = 1 - \frac{1}{n} \quad p^n = \left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0.368$$
- Trainiere mit der Stichprobe und verwende alle Datensätze, die nicht in der Stichprobe vorkommen als Testdatensätze

## 0.632-Bootstrap (2)

- Trainiere mit der Stichprobe und verwende alle Datensätze, die nicht in der Stichprobe vorkommen als Testdatensätze
- Ermittle die Fehlerrate als Kombination aus dem Fehler auf den Test- und Trainingsdaten

$$\text{Fehlerrate} = 0.632 \text{Fehlerrate}_{\text{Testdatensätze}} + 0.368 \text{Fehlerrate}_{\text{Trainingsdatensätze}}$$

## Kennzahlen eines Klassifikators

Sensitivität =  $P(\text{Test positiv} \mid \text{krank}) = A/(A+C)$

Spezifität =  $P(\text{Test negativ} \mid \text{gesund}) = D/(B+D)$

Wie viele gefundene Ergebnisse sind richtig?

Precision =  $A/(A+B)$

Wie viele richtige Dokumente kann ich finden?

Recall =  $A/(A+C)$

Klassifikationsfehler

AUC (area under ROC curve)

	Krankheit vorhanden	Krankheit nicht Vorhanden
Test positiv	Richtig positiv (A)	Falsch positiv (B)
Test negativ	Falsch negativ (C)	Richtig negativ (D)

## Analytische Bewertung eines Klassifikators

---

- AIC
  - Akaike information criterion
- MDL
  - Minimum description length
- VC-Dimension

## Zusammenfassung

---

- Reklassifikation, empirischer Fehler
  - Test und Trainingsdaten sind identisch
- Holdout, Subdivision
  - Test und Trainingsdaten sind getrennt
- Kreuzvalidierung (cross validation)
  - Aus allen Trainingsdaten werden zufällig einige Datensätze entfernt und als Testdaten verwendet
  - Der Prozess wird n mal wiederholt
  - Spezialfall: leave-one-out-crossvalidation

## Kosten für eine Fehlklassifikation

---

- Mammographie
- Weiterentwicklung eines Targets
- Kreditvergabe
- Werbebriefe

## Motivation lift chart

---

- Werbebriefe an alle Kunden: 0.1 % Resonanz
- Werbebriefe an 5% der Kunden: 10% Resonanz
- Werbebriefe an 20 Kunden: 100% Resonanz
  
- An wie viele Kunden sollen Brief versandt werden?
  - Kosten pro Werbebrief
  - Gewinn pro Resonanz des Kunden

## Steigerungsdiagramm (Lift chart)

---

- Ordne alle Klassifikationsergebnisse

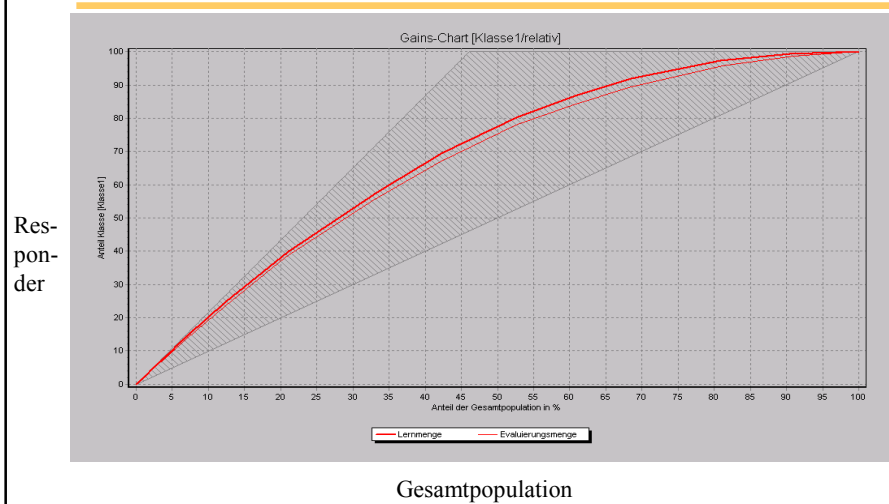
Rang	Wahrscheinlichkeit für Klasse 1	Klassifikation	Y-Achse/X-Achse
1	0.95	Richtig	1/1
2	0.9	Falsch	1/2
3	0.87	Richtig	2/3
4	0.85	Falsch	2/4

## Steigerungsdiagramm (Lift chart)

---

- Anzahl klassifizierter Datensätze zur Anzahl der richtig Klassifizierten
- Lift-Faktor: Antwort-Quote der Stichprobe im Vergleich zur globalen Antwort-Quote
- Marketing

## Lift chart



DKFZ Heidelberg

43

Falk Schubert

## ROC-Kurve

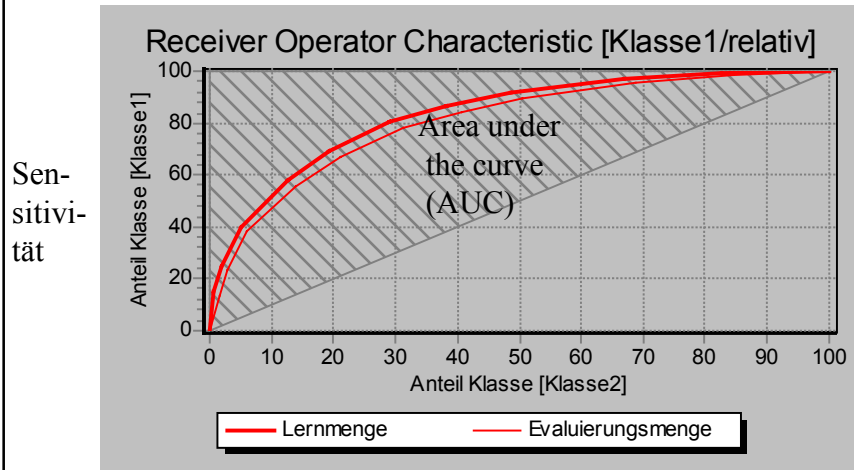
- Receiver-Operator-Characteristics
- 1-Spezifität zur Sensitivität
- Fläche unter der Kurve Maß für die Güte des Klassifikators

DKFZ Heidelberg

44

Falk Schubert

# ROC



DKFZ Heidelberg

1-Spezifität  
45

Falk Schubert

DKFZ Heidelberg

Danke!