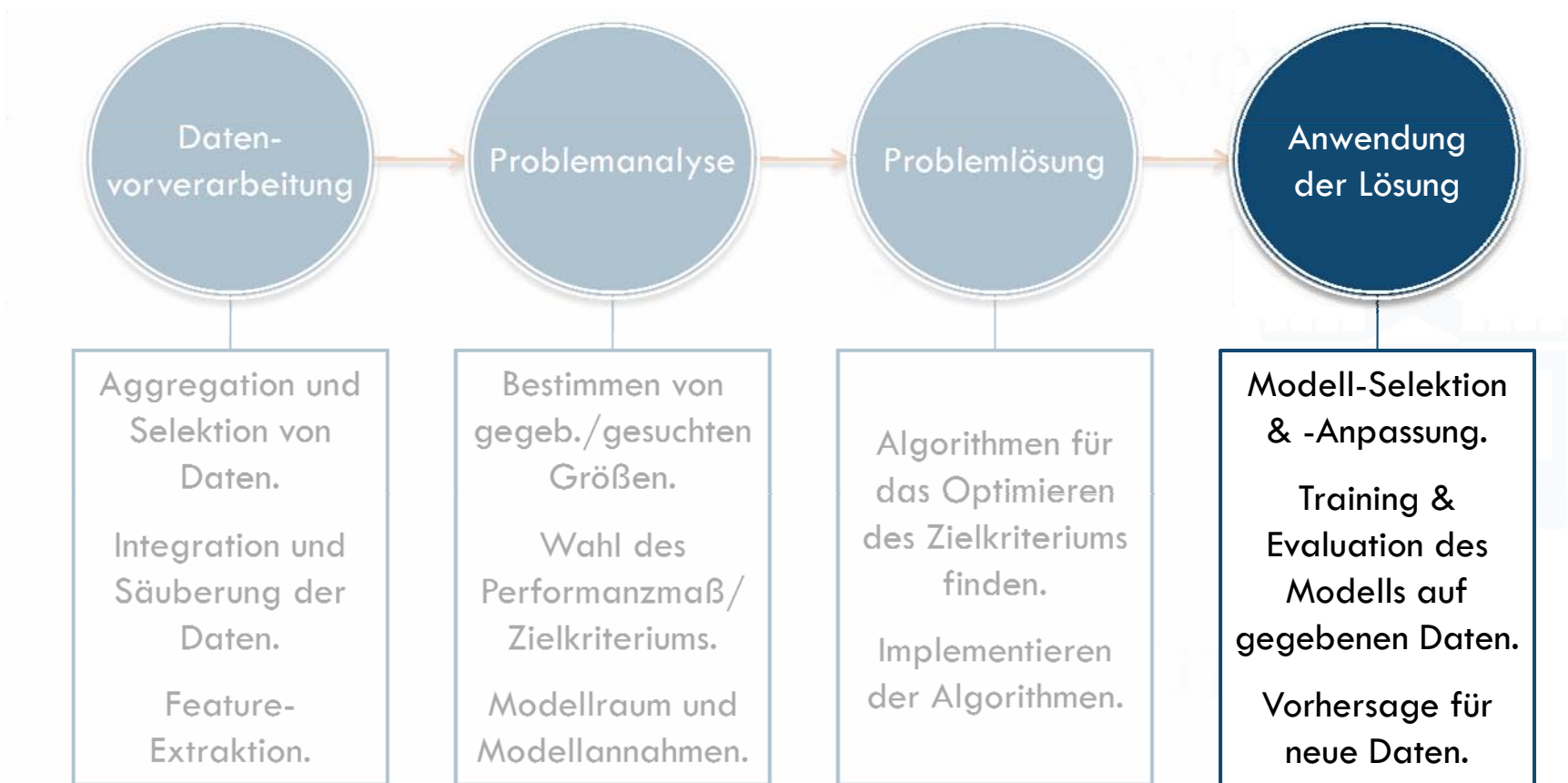


# INTELLIGENTE DATENANALYSE IN MATLAB

Evaluation & Exploitation von Modellen

# Überblick

## □ Schritte der Datenanalyse:



# Überblick

- Evaluation von Lernverfahren.
- Selektion und Anpassung von Modellen.
- Evaluation von Klassifikatoren.
- Exploitation von Modellen.

# Evaluation von Lernverfahren

- Ziel: Qualitätsbewertung der Modelle eines Lernverfahrens.
  - Nachdem wir Problem analysiert haben und Verfahren identifiziert & implementiert haben.
  
- Qualität eines Modells: Wie gut sind die Vorhersagen des Modells?
  - Was genau heißt „gut“?
  - Wie berechnet/schätzt man die Genauigkeit der Vorhersagen (für zukünftige Daten)?

# Evaluation von Lernverfahren

## Problemstellung

### □ Gegeben:

□ Repräsentative Evaluierungsdaten  $E$   
mit bekannter Zielgröße.

□ Bewertungsmaß (Verlustfunktion)  
welche Qualität einer Vorhersage  
misst, z.B.

Muss nicht identisch sein  
zur Verlustfunktion des  
Lernverfahrens

■ Klassifikation: Anzahl falsch klassifizierter Beispiele (Fehlerrate).

$$l(y^{prediction}, y) = [y^{prediction} \neq y]$$

■ Regression: Mittlerer quadratischer Fehler.

$$l(y^{prediction}, y) = (y^{prediction} - y)^2$$

■ Ranking: Mittlerer Abstand zw. echter und vorhergesagter Position.

# Evaluation von Lernverfahren

## Problemstellung

- Eingabe: Lernverfahren welches ein Modell  $h$  ausgibt.
- Ziel: Bewertung der mittleren Qualität des Lernverfahrens.
  - Theoretischer Mittelwert des Verlusts auf der Testverteilung.

$$R_{theo} = E[l(h(X), Y)] = \sum_{x,y} p(x, y)l(h(x), y)$$

- Aber: Testverteilung  $p(X, Y)$  unbekannt!
- Evaluierungsdaten  $E = \{(x_1, y_1), \dots, (x_n, y_n)\}$  sind repräsentativ aus  $p(X, Y)$  gezogen  $\Rightarrow$  theoretischen Mittelwert durch empirischen Mittelwert (empirisches Risiko) schätzen:

$$R_{emp} = \frac{1}{n} \sum_{i=1}^n l(h(x_i), y_i)$$

# Evaluation von Lernverfahren

## Problemstellung

- Welche Daten für Evaluation verwenden:
  - ▣ Daten auf welchen das Modell trainiert wurde?  
**Nein!** Empirischer Verlust auf diesen Daten meist 0.
  - ▣ Daten auf welche das Modell angewendet werden soll?  
**Nein!** Zielgröße für diese Daten unbekannt.
- Idee:
  - ▣ Gelabelte Trainingsdaten aufteilen in
    - *Lerndaten* zum Lernen eines Modells, und
    - *Evaluierungsdaten* zum Evaluieren des Modells.

# Evaluation von Lernverfahren

## Aufteilung der Trainingsdaten: Holdout Validation

- Gegeben: Trainingsdaten  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ .
- Aufteilen der Daten in Lerndaten  $L = \{(x_1, y_1), \dots, (x_k, y_k)\}$  und Evaluierungsdaten  $E = \{(x_{k+1}, y_{k+1}), \dots, (x_n, y_n)\}$ .
- Lerne Modell  $h'$  auf Daten  $L$  und bestimme empirisches Risiko auf Daten  $E$ :  $R_{emp}(h') = \frac{1}{n-k} \sum_{i=k+1}^n l(h'(x_i), y_i)$
- Lerne Modell  $h$  auf Daten  $D$ .
- Ausgabe: Modell  $h$  mit Risiko-Schätzer  $\hat{R}_{emp}(h) = R_{emp}(h')$ .

Pessimistische Schätzung



# Evaluation von Lernverfahren

## Aufteilung der Trainingsdaten: Cross Validation

- Gegeben: Trainingsdaten  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ .
- Aufteilen der Daten in  $p$  Blöcke  $D_i = \{(x_{i_1}, y_{i_1}), \dots, (x_{i_k}, y_{i_k})\}$  mit  $D = \bigcup_i D_i$  und  $D_i \cap D_j = \emptyset$  für 2 verschiedene Blöcke.
- Wiederhole für  $i = 1 \dots p$ 
  - ▣ Trainiere Modell  $h_i$  auf Daten  $D \setminus D_i$ .
  - ▣ Berechne empirisches Risiko auf  $D_i$ :  $R_{emp}(h_i) = \frac{1}{k} \sum_{j=1}^k l(h_i(x_{i_j}), y_{i_j})$
- Lerne Modell  $h$  auf Daten  $D$ .
- Ausgabe: Modell  $h$  mit mittlerem Risiko  $\hat{R}_{emp}(h) = \frac{1}{p} \sum_{i=1}^p R_{emp}(h_i)$ .

# Evaluation von Lernverfahren

## Aufteilung der Trainingsdaten: Leave-One-Out Validation

- Gegeben: Trainingsdaten  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ .
- Spezialfall von Cross Validation mit  $D_i = (x_i, y_i)$ .
- Wiederhole für  $i = 1 \dots n$ 
  - ▣ Trainiere Modell  $h_i$  auf Daten  $D \setminus (x_i, y_i)$ .
  - ▣ Berechne empirisches Risiko für  $(x_i, y_i)$ :  $R_{emp}(h_i) = l(h_i(x_i), y_i)$
- Lerne Modell  $h$  auf Daten  $D$ .
- Ausgabe: Modell  $h$  mit Loo-Fehler  $\hat{R}_{emp}(h) = \frac{1}{n} \sum_{i=1}^n R_{emp}(h_i)$ .
  - ▣ I.d.R. aufwendig zu berechnen.
  - ▣ Für einige Probleme existiert analytische Lösung für Loo-Fehler.

# Evaluation von Lernverfahren

## Signifikanz des empirischen Risikos

- Wie gut ist der Schätzer  $\hat{R}_{emp}(h)$  für das echte Risiko  $R_{theo}(h)$ ?
- Idee:  $m$ -malige Validation ergibt  $m$  Schätzwerte für empirisches Risiko mit Mittelwert  $\mu_R$ .

- ▣ Standardfehler (Standardabw. des Schätzers):  $\sigma_R = \sqrt{\frac{\mu_R(1-\mu_R)}{m-1}}$

- ▣ Test der Hypothese  $|R_{theo}(h) - \hat{R}_{emp}(h)| \leq \varepsilon$ :

$$\begin{aligned}
 p\left(\left|R_{theo}(h) - \hat{R}_{emp}(h)\right| \leq \varepsilon\right) &= 1 - \left(p\left(R_{theo}(h) - \hat{R}_{emp}(h) > \varepsilon\right) + p\left(\hat{R}_{emp}(h) - R_{theo}(h) > \varepsilon\right)\right) \\
 &\approx 1 - 2\Phi\left(-\varepsilon\sigma_R^{-2}\right)
 \end{aligned}$$

Dichtefunktion der Normalverteilung

# Evaluation von Lernverfahren

## Signifikanz des empirischen Risikos

- Test der Hypothese  $|R_{theo}(h) - \hat{R}_{emp}(h)| \leq \varepsilon$  mit Signifikanzniveau 5% (*signifikantes Ereignis*).
- Beispiel: 10-malige Wiederholung einer Leave-One-Out-Validation (auf 10 verschiedenen Datensätzen).
  - 10 Schätzwerte mit Mittelwert  $\mu_R = 8\% \Rightarrow \sigma_R = 0,09$ .
  - Gesucht ist  $\varepsilon$  mit Konfidenzintervall  $1 - \delta$  und  $\delta = 5\%$ :

$$\begin{aligned}
 p\left(|R_{theo}(h) - \hat{R}_{emp}(h)| \leq \varepsilon\right) &\geq 0,950 && \Phi(z) \geq 0,975 \\
 1 - 2\Phi(-0,09^{-2} \cdot \varepsilon) &\geq 0,950 &\Rightarrow & z = 0,835 \\
 \Phi(123,3 \cdot \varepsilon) &\geq 0,975 && \varepsilon = \frac{z}{123,3} = 0,68\% && \Rightarrow \mu_R = 8,0 \pm 0,68\%
 \end{aligned}$$

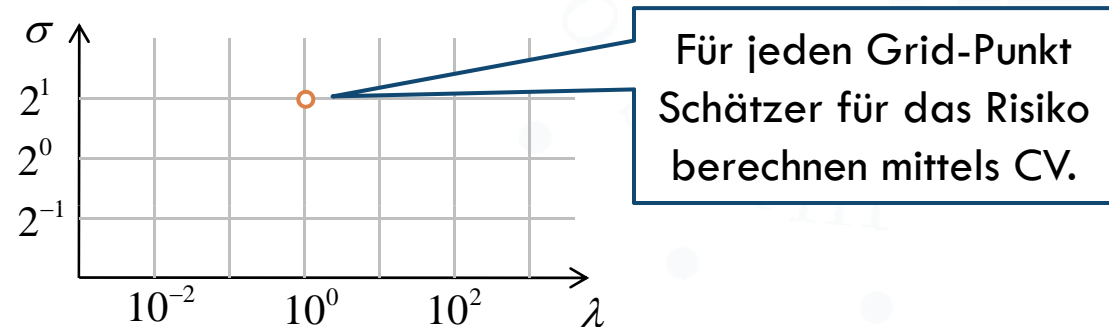
# Selektion und Anpassung von Modellen

- Ziel: Hohe Qualität des Modells durch Selektion/Anpassung des Modells bzw. Lernverfahrens.
- Anpassen von
  - Modellkomponenten (z.B. Verlustfunktion/Regularisierung, Splitting-Kriterium).
  - Parameter des Lernverfahrens (z.B. maximale Anzahl Iterationen).
  - Parameter der Verlustfunktion (z.B. Klassen-Kosten).
  - Parameter des Regularisierers (z.B.  $\lambda$  des  $\Omega_2$ -Regularisierers).
  - Parameter der Daten-Transformation bzw. des Kernels (z.B.  $\sigma$  des RBF-Kernels).

# Selektion und Anpassung von Modellen

## Grid-Suche

- Idee:
  - Stichprobenartig aus der Menge aller möglichen Parameter bzw. Parameterkombinationen ziehen.
  - Für jede gezogene Kombination mittels Cross-Validation (CV) Schätzer für  $R_{theo}(h)$  bestimmen.
  - Parameter wählen mit minimalem Risiko.
- Beispiel für Parameter-Auswahl: *Grid-Suche*



# Selektion und Anpassung von Modellen

## Aufteilung der Lerndaten



- Welche Daten für Modell-Anpassung verwenden:
  - ▣ Daten auf welchen das Modell evaluiert wird?  
**Nein!** Evaluierung des Modells wäre zu optimistisch.
- Idee:
  - ▣ Lerndaten aufteilen in Daten für ...
    - *Learning*: zum Lernen eines Modells mit festen Parametern und
    - *Tuning*: zum Anpassen der Modellparameter.
  - ▣ Art der Aufteilung:
    - Holdout-Validation.
    - Cross-Validation.
    - Loo-Validation.

# Selektion und Anpassung von Modellen

## Aufteilung der Lerndaten



- Beispiel: Geschachtelte Cross-Validation.
  - Aufteilen der Trainingsdaten  $D$  in  $p$  Blöcke  $D_i$ .
  - Wiederhole für  $i = 1 \dots p$ 
    - Aufteilen der Lerndaten  $L = D \setminus D_i$  in  $q$  Blöcke  $L_j$ .
    - Wiederhole für alle Modell-Parameterkombinationen
      - Wiederhole für  $j = 1 \dots q$ 
        - Trainiere für aktuelle Parameterkombination ein Modell auf  $L \setminus L_j$ .
        - Berechne empirisches Risiko auf  $L_j$ .
      - Bestimme mittleres empirisches Risiko für aktuelle Parameterkombination.
    - Trainiere für beste Parameterkombination Modell  $h_i$  auf  $D \setminus D_i$ .
    - Berechne empirisches Risiko auf  $D_i$ .
  - Trainiere für beste Parameterkombination Modell  $h$  auf  $D$ .



# Evaluation von Klassifikatoren

- Ziel: Bewertung eines konkreten Modells für binäre Klassifikation.
  - Nachdem wir Problem analysiert haben, Verfahren identifiziert & implementiert haben, und Klassifikations-Modell (Klassifikator) trainiert haben.
  
- Qualität eines Klassifikators:
  - Precision/Recall-Analyse.
  - ROC-Analyse.

# Evaluation von Klassifikatoren

## Definitionen (für binäre Klassifikation)

- Entscheidungsfunktion: Ordnet einer Eingabe  $\mathbf{x}$  einen numerischen Wert zu,  $f : \mathbb{R}^m \rightarrow \mathbb{R}$ .
  - Beispiel:  $f(\mathbf{x}) = \mathbf{x}^T \mathbf{w}$
- Klassifikationsfunktion: Ordnet einem Entscheidungsfunktionswert  $f(\mathbf{x})$  ein Klassenlabel zu,  $g : \mathbb{R} \rightarrow Y$ .
  - Beispiel:  $g(f(\mathbf{x})) = \text{sign}(f(\mathbf{x}) + \theta)$
- Kontingenztafel:

Klassifikations-Schwellwert

	Tatsächlich positiv	Tatsächlich negativ
Positiv vorhergesagt	TP (true positives)	FP (false positives)
Negativ vorhergesagt	FN (false negatives)	TN (true negatives)

# Evaluation von Klassifikatoren

## Definitionen (für binäre Klassifikation)

- Beispiel HIV-Erkrankungen in Deutschland (2008):
  - ▣ In Deutschland leben 82.099.232 Menschen.
  - ▣ Davon sind 63.554 Menschen an HIV erkrankt.
  - ▣ Ein HIV-Test ergab (hochgerechnet auf alle Menschen):

	Tatsächlich positiv	Tatsächlich negativ	Summe
Positiv vorhergesagt	63.487	114.276	177.763
Negativ vorhergesagt	67	81.921.402	81.921.469
Summe	63.554	82.035.678	82.099.232

False Negatives:  
fälschlicherweise als  
HIV-negativ klassifiziert

False Positives:  
fälschlicherweise als  
HIV-positiv klassifiziert

# Evaluation von Klassifikatoren

## Qualität eines Klassifikators

- **Gegeben:**
  - Repräsentative Evaluierungsdaten  $E$  mit bekannter Zielgröße.
  - Entscheidungs- und Klassifikationsfunktion.
- **Gesucht:**
  - Bewertung der Entscheidungsfunktion.
    - Beispiele: Precision/Recall-Kurve, ROC-Kurve.
  - Bewertung der Klassifikationsfunktion (Entscheidungsfunktion für einen konkreten Schwellwert).
    - Beispiele: Fehlerrate, F-Maß.

# Evaluation von Klassifikatoren

## Qualität eines Klassifikators

- Für jeden Klassifikations-Schwellwert ergibt sich eine Kontingenztafel, d.h. Werte für  $TP$ ,  $FP$ ,  $TN$  und  $FN$ .
- Unterschiedliche Bewertungsmaße für einen Klassifikator (für einen konkreten Schwellwert):

■ Trefferquote (Recall):  $\frac{TP}{TP + FN} = \frac{63.487}{63.487 + 67} = 99,89\%$

■ Genauigkeit (Precision):  $\frac{TP}{TP + FP} = \frac{63.487}{63.487 + 114.276} = 35,71\%$

■ Ausfallquote (Fallout):  $\frac{FP}{TN + FP} = \frac{114.276}{81.921.402 + 114.276} = 0,14\%$

# Evaluation von Klassifikatoren

## Qualität eines Klassifikators

	Tatsächlich positiv	Tatsächlich negativ	Summe
Positiv vorhergesagt	63.487	114.276	177.763
Negativ vorhergesagt	67	81.921.402	81.921.469
Summe	63.554	82.035.678	82.099.232

- Trefferquote (Recall):  $\frac{TP}{TP + FN} = \frac{63.487}{63.487 + 67} = 99,89\%$
- Genauigkeit (Precision):  $\frac{TP}{TP + FP} = \frac{63.487}{63.487 + 114.276} = 35,71\%$
- Ausfallquote (Fallout):  $\frac{FP}{TN + FP} = \frac{114.276}{81.921.402 + 114.276} = 0,14\%$

# Evaluation von Klassifikatoren

## Recall versus Precision

- Kombinierte Bewertungsmaße aus Recall und Precision:
  - Sensitivität (Sensitivity): Recall bzgl. positiver Beispiele.
  - Spezifität (Specificity): Recall bzgl. negativer Beispiele.
  - F-Maß (F-score): Harmonisches Mittel aus Precision & Recall.

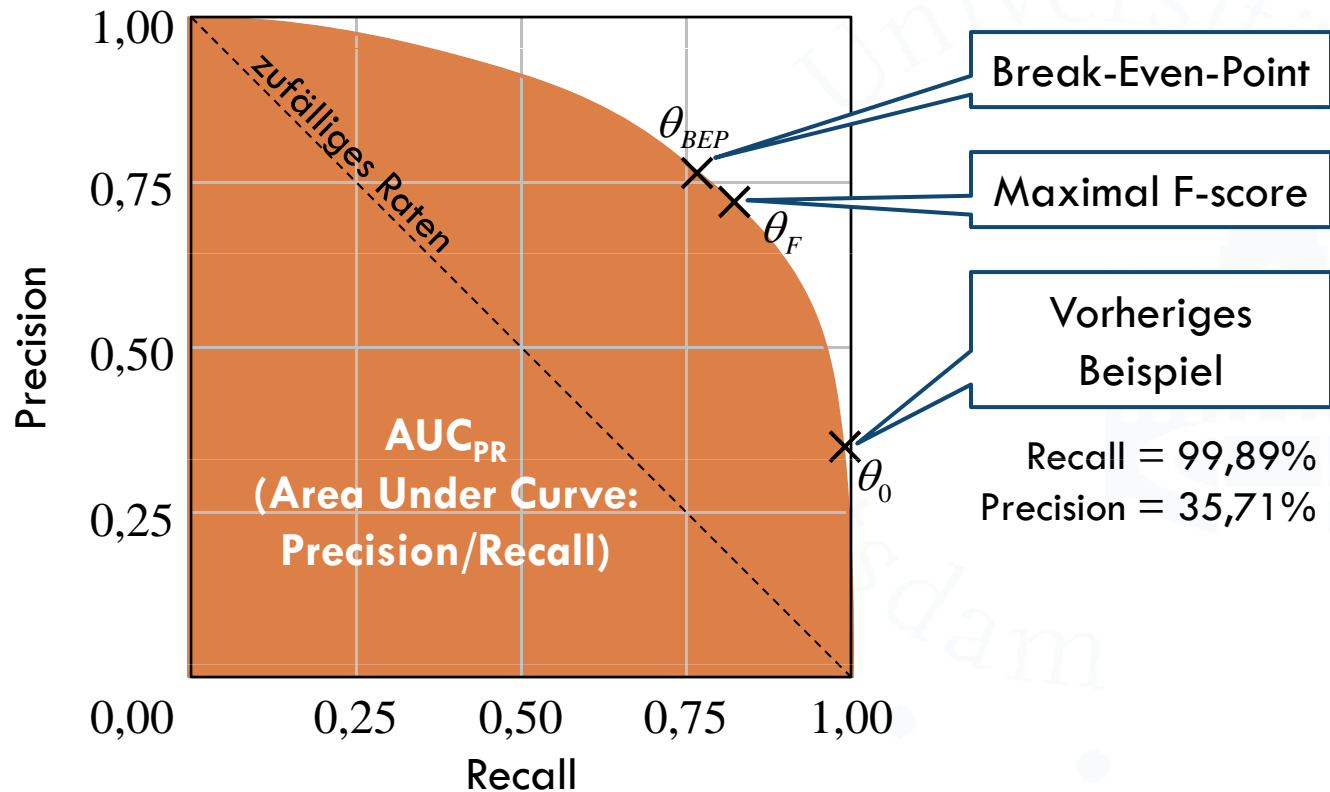
$$\text{F-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \cdot \frac{35,71\% \cdot 99,89\%}{35,71\% + 99,89\%} = 52,61\%$$

- Spezielle Schwellwerte  $\theta$ :
  - Gewinnschwelle (Break-Even-Point): Schwellwert für welchen Precision = Recall.
  - F-Schwellwert (Maximal F-score): Schwellwert für welchen F-score maximal ist.

# Evaluation von Klassifikatoren

## Recall versus Precision

- Precision/Recall-Kurve: Precision vs. Recall für unterschiedliche Schwellwerte  $\theta$ .





# Evaluation von Klassifikatoren

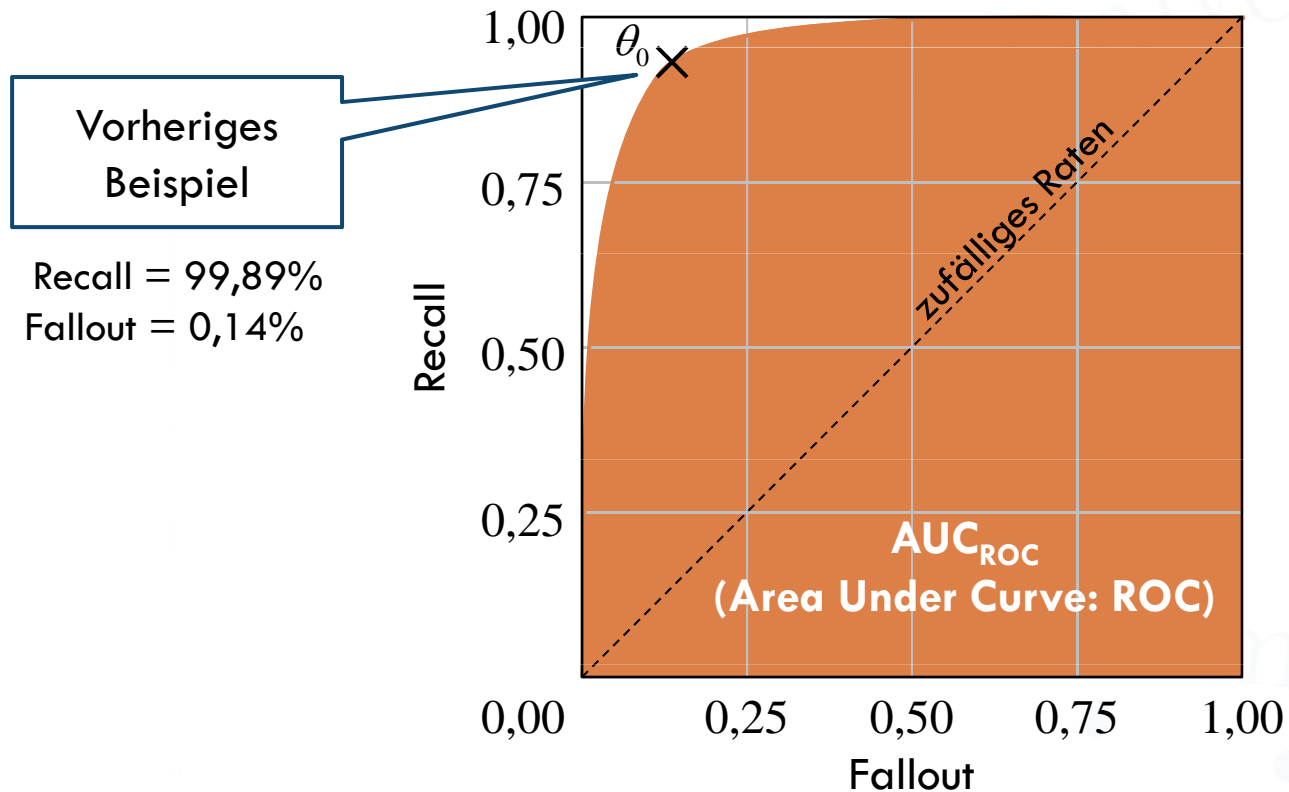
## Recall versus Fallout

- Receiver-Operating-Characteristic (ROC):
  - Bewertung der Entscheidungsfunktion unabhängig vom Schwellwert durch Fallout statt Precision.
  - Großer Schwellwert: Mehr positive Beispiel falsch klassifiziert.
  - Kleiner Schwellwert: Mehr negative Beispiel falsch klassifiziert.
- Fläche unter der ROC-Kurve ( $AUC_{ROC}$ ) bewertet Entscheidungsfunktion.
  - Analog zur Fläche unter Precision/Recall-Kurve.

# Evaluation von Klassifikatoren

## Recall versus Fallout

- ROC-Kurve bzw. Recall/Fallout-Kurve: Recall (True Positives Rate) vs. Fallout (False Positives Rate).



# Evaluation von Klassifikatoren

## Recall versus Fallout

### □ Algorithmus zur Bestimmung des $AUC_{ROC}$ -Wertes.

$AUC_{ROC}(f, y)$

Sortiere Paare  $(f_i, y_i)$   
aufsteigend nach  $f_i$

Setze  $TN = 0, FN = 0, AUC = 0$

FOR  $i = 1 \dots n$

IF  $y_i > 0$  THEN

$FN = FN + 1$

$AUC = AUC + TN$

ELSE

$TN = TN + 1$

$AUC = AUC / (FN * TN)$

RETURN  $AUC$

**f** Vektor mit  $n$  Entscheidungsfunktionswerten  
**y** Vektor mit zugehörigen Klassenlabels

# Exploitation von Modellen

- Anwenden von Modellen in der Praxis:
  - Einstellen von Modellparametern nach dem Lernen (z.B. Schwellwerte, Default-Klasse).
  - Kombination mehrerer gelernter Modelle (z.B. Verwendung mehrerer Spam-Filter).
  - Integration des Modells in bestehende Softwarearchitektur.
  - Monitoren der Qualität (Verteilung der Eingabedaten ändert sich oft über die Zeit  $\Rightarrow$  Qualität verringert sich).

# Zusammenfassung

- Qualität von Lernverfahren/Modellen messen...
  - Auf Evaluierungsdaten; nicht auf Trainingsdaten!
  - Signifikanz des Ergebnisses prüfen.
- Modell-Selektion/-Anpassung...
  - Auf Tuningdaten; nicht auf Evaluierungsdaten!
  - Modellparameter z.B. durch Grid-Suche + Cross-Validation.
- Bewertung eines Klassifikators durch Recall, Precision, Fallout, F-Maß usw.
- Bewertung einer Entscheidungsfunktion durch (Fläche unter der) Precision/Recall-Kurve oder ROC-Kurve.