Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Hypothesenbewertung

Christoph Sawade/Niels Landwehr
Silvia Makowski
Tobias Scheffer

Überblick

- Hypothesenbewertung, Risikoschätzung
- Anwendungen/Beispiele
- Konfidenzintervalle
- Roc-Analyse

Überblick

- Hypothesenbewertung, Risikoschätzung
- Anwendungen/Beispiele
- Konfidenzintervalle
- Roc-Analyse

Lernen und Vorhersage

- Klassifikation, Regression: Lernproblem
  - Eingabe: Trainingsdaten $L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m) \rangle$
  - Ausgabe: Hypothese (Modell) $f : X \to Y$

- Modell wird verwendet, um Vorhersagen für neue Testbeispiele zu treffen

$$f(\mathbf{x}) = ? \in \mathcal{Y} \qquad \mathbf{x} \in \mathcal{X}$$ Testbeispiel

- Verschiedene Verfahren kennengelernt
  - Lineare Modelle, Entscheidungsbäume, Naive Bayes…

Hypothesenbewertung

- Frage: Nachdem wir Verfahren implementiert, Hypothese trainiert haben etc.: wie gut sind die Vorhersagen?
  - Was genau heißt „gut“?
  - Wie berechnet / misst / schätzt man es?

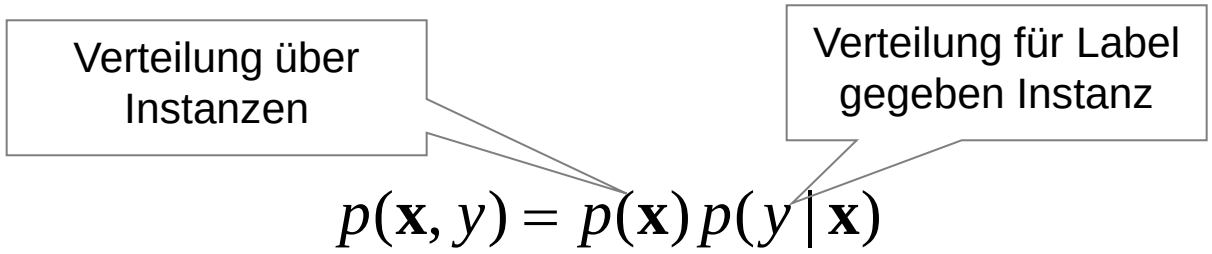
- „Gute Vorhersagen“: in der Zukunft, wenn die gelernte Hypothese eingesetzt wird (auf Testdaten)

- „Hypothesenbewertung“: Abschätzung der Genauigkeit/Güte von Vorhersagen gelernter Modelle

Hypothesenbewertung

- Um Hypothesenbewertung formal zu untersuchen, müssen wir Annahmen über die Eigenschaften von Trainings- und Testdaten machen

- Zentrale Annahme: dem Lernproblem liegt eine (unbekannte) Verteilung $p(\mathbf{x},y)$ zugrunde



- Beispiel Spam-Filterung
  - $p(\mathbf{x})$ Wahrscheinlichkeit dass wir Email $\mathbf{x}$ sehen
  - $p(y|\mathbf{x})$ Wahrscheinlichkeit dass Mensch Email $\mathbf{x}$ als $y \in \{\text{Spam/Ok}\}$ klassifiziert

Hypothesenbewertung

- „i.i.d.“-Annahme: Beispiele sind „independent and identically distributed“.
  - Trainingsdaten werden unabhängig aus der Verteilung $p(\mathbf{x},y)$ gezogen:

$$(\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$$ Trainingsbeispiele

  - Testdaten ebenfalls i.i.d. aus dieser Verteilung gezogen

$$(\mathbf{x}, y) \sim p(\mathbf{x}, y)$$ Testbeispiel (Anwendung des Modells)

  - Immer realistisch?

- Wir nehmen im Folgenden immer i.i.d. Daten an

Verlustfunktionen

- Wir haben Annahmen darüber gemacht, welche Instanzen $(\mathbf{x},y)$ auftauchen werden
- Testbeispiel $(\mathbf{x},y)$, Hypothese sagt $f(\mathbf{x})$.
- Verlustfunktion definiert, wie schlecht das ist:

$\ell(y, f(\mathbf{x}))$ Verlust der Vorhersage $f(\mathbf{x})$ auf Instanz $(\mathbf{x},y)$

  - Nicht-negativ: $\forall y, y' : \ell(y, y') \geq 0$
  - Problem-spezifisch, gegeben.
- Verlustfunktionen für Klassifikation
  - Zero-one loss: $\ell(y, y') = 0, \ wenn \ \ y = y'; \ \ 1, \ sonst$
  - Klassenabhängige Kostenmatrix
- Verlustfunktionen für Regression
  - Squared loss: $\ell(y, y') = (y - y')^2$

Hypothesenbewertung: Risiko

- Zentraler Begriff der Hypothesenbewertung: Risiko
- Risiko einer Hypothese: erwarteter Verlust für eine neue Instanz

$(\mathbf{x}, y) \sim p(\mathbf{x}, y)$ Testbeispiel $\mathbf{x}$ mit Label y (Zufallsvariable)

$\ell(y, f(\mathbf{x}))$ Verlust auf Testbeispiel (Zufallsvariable)

$$R(f) = E[\ell(y, f(\mathbf{x}))] = \int \ell(y, f(\mathbf{x}))\, p(\mathbf{x}, y)\, d\mathbf{x}\, d$$

  - Für Zero-One-Loss heisst Risiko auch Fehlerrate.
  - Für Squared Loss heisst Risiko auch Mean Squared Error.

- Hauptziel der Hypothesenbewertung: Risikobestimmung
- Nicht möglich, das Risiko exakt zu bestimmen, weil $p(\mathbf{x},y)$ unbekannt ist → Schätzproblem

Einschub: Begriff „Schätzer“

- Wir wollen das Risiko einer Hypothese aus Daten schätzen
- Formalisierung: ein **Schätzer** ist ein Verfahren, das Beobachtungen L auf einen Schätzwert abbildet.
- Beispiel Münzwurf:
  - Beobachtung N_k, N_z: wie oft Kopf/Zahl gesehen
  - schätze Münzparameter θ (Wahrscheinlichkeit, dass Kopf fällt)
- Notation: Schätzer für (unbekannten) Wert θ wird mit $\hat{\theta}$ bezeichnet

Hypothesenbewertung:Risikoschätzung

- Risikoschätzung aus Daten
- Wenn aus $p(\mathbf{x},y)$ gezogene Daten

$$T = \left\langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \right\rangle \qquad (\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$$

gegeben sind, kann das Risiko geschätzt werden:

$$\hat{R}(f) = \frac{1}{m} \sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j)) \qquad \text{"Empirisches Risiko"}$$

- Wichtig: Wo kommt T her?
  - Trainingsdaten (T=L)?
  - Verfügbare Daten in disjunkte L und T aufteilen.
  - Cross-Validation.

Bias eines Schätzers

- Schätzer

$$\hat{R}(f) = \frac{1}{m} \sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j))$$

- Schätzer ist Zufallsvariable:
  - Instanzen in T zufällig gezogen

$$(\mathbf{x}_j, y_j) \sim p(\mathbf{x}, y) \qquad \text{Welche } (\mathbf{x}_j, y_j) \text{ werden gezogen?}$$

  - Wert des Schätzers hängt von zufällig gezogenen Instanzen ab, ist also Produkt eines Zufallsprozesses

- Schätzer hat einen Erwartungswert
- Schätzer ist erwartungstreu, genau dann wenn:
  - Erwartungswert des empirischen Risikos = echtes Risiko.

Bias eines Schätzers

- Schätzer $\hat{R}(f)$ ist erwartungstreu, genau dann wenn:
  - $E[\hat{R}(f)] = R(f)$
- Ansonsten hat $\hat{R}(f)$ einen Bias:
  - $Bias = E[\hat{R}(f)] - R(f)$
- Schätzer ist optimistisch, wenn
  - *Bias* < 0.
- Schätzer ist pessimistisch, wenn
  - *Bias* > 0.
- Schätzer ist erwartungstreu, wenn
  - *Bias* = 0.

Varianz eines Schätzers

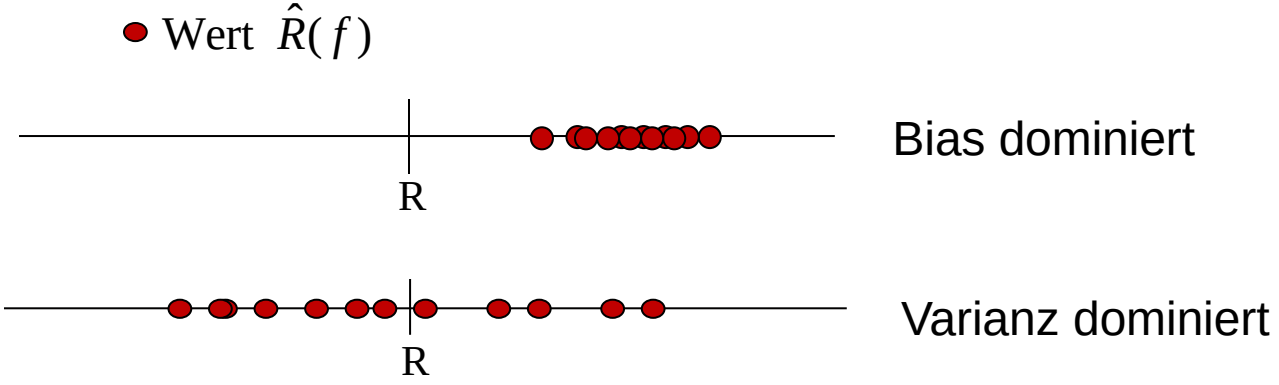
- Schätzer $\hat{R}(f)$ hat eine Varianz

$$Var[\hat{R}(f)] = E[\hat{R}(f)^2] - E[\hat{R}(f)]^2$$

- Je größer die Stichprobe ist, die zum Schätzen verwendet wird, desto geringer ist die Varianz.

- Varianz vs. Bias:
  - Hohe Varianz: großer „Zufallsanteil“ bei der Bestimmung des empirischen Risikos.
  - Großer Bias: systematischer Fehler bei der Bestimmung des empirischen Risikos.

Varianz eines Schätzers



- Erwarteter (quadratischer) Fehler des Schätzers:

$$E[(\hat{R}(f)-R)^2]$$

- Lässt sich zerlegen in Bias und Varianz

$$\begin{aligned}
E[(\hat{R}(f)-R)^2] &= E[\hat{R}(f)^2 - 2R\hat{R}(f) + R^2] \\
&= E[\hat{R}(f)^2] - 2RE[\hat{R}(f)] + R^2 \\
&= E[\hat{R}(f)]^2 - 2RE[\hat{R}(f)] + R^2 + E[\hat{R}(f)^2] - E[\hat{R}(f)]^2 \\
&= (E[\hat{R}(f)]-R)^2 + Var[\hat{R}(f)] \\
&= Bias[\hat{R}(f)]^2 + Var[\hat{R}(f)]
\end{aligned}$$

Verschiebungssatz

Risikoschätzer auf den Trainingsdaten

- Welche Menge T verwenden?
- 1. Versuch: Trainingsdaten L

- Hypothese f, trainiert auf $L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m) \rangle$

- Empirisches Risiko gemessen auf Trainingsdaten

$$\hat{R}_L(f) = \frac{1}{m} \sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j))$$

- Ist dieses empirische Risiko ein
  - erwartungstreuer
  - optimistischer
  - pessimistischer
- Schätzer für das echte Risiko und warum?

Risikoschätzer auf den Trainingsdaten

- Empirisches Risiko auf den Trainingsdaten ist ein optimistischer Schätzer des echten Risikos

- Empirisches Risiko *aller möglichen* Hypothesen für ein festes L?
  - Aufgrund von Zufallseffekten gilt für einige Hypothesen f, dass $\hat{R}_L(f) < R(f)$ und für andere Hypothesen f, dass $\hat{R}_L(f) > R(f)$.
  - Lernalgorithmus wählt eine Hypothese f_L mit kleinem empirischen Risiko $\hat{R}_L(f_L)$
  - Wahrscheinlich, dass $\hat{R}_L(f_L) < R(f_L)$ (Risiko unterschätzt)

Risikoschätzer auf den Trainingsdaten

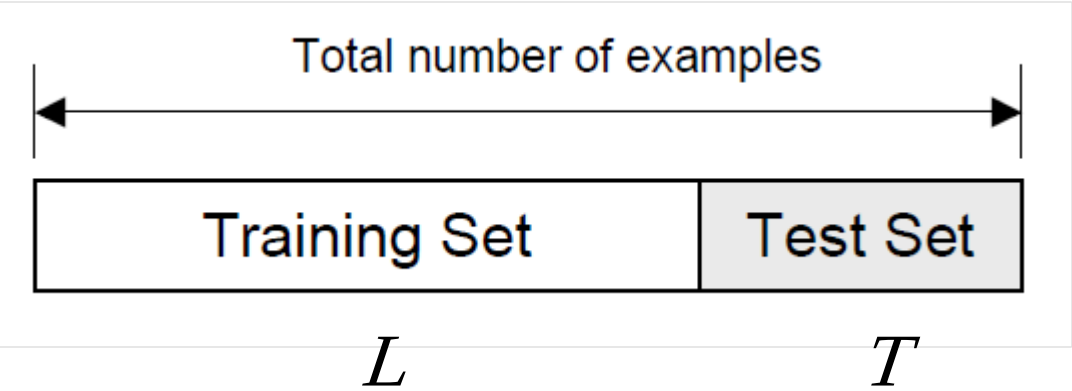
- Empirisches Risiko der vom Lerner gewählten Hypothese auf den Trainingsdaten („Trainingsfehler“) ist ein optimistischer Schätzer des echten Risikos:

$$E[\hat{R}_L(f_L)] < R(f_L)$$

- Problem ist die Abhängigkeit von gewählter Hypothese und zur Fehlerschätzung verwendeten Daten
- Ansatz: Testdaten verwenden, die von den Trainingsdaten unabhängig sind.

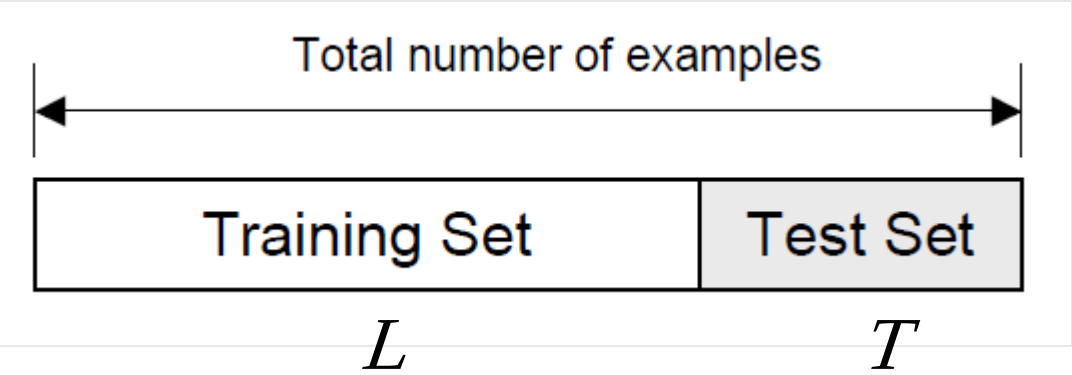
Holdout-Testing

- Idee: Fehlerschätzung auf unabhängigen Testdaten
- Gegeben: Daten $D = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_d, y_d) \rangle$
- Teile Daten auf in
  - Trainingsdaten $L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m) \rangle$ und
  - Testdaten $T = \langle (\mathbf{x}_{m+1}, y_{m+1}), ..., (\mathbf{x}_d, y_d) \rangle$



Holdout-Testing

- Starte Lernalgorithmus mit Daten L, gewinne so Hypothese f_L.
- Ermittle empirisches Risiko $\hat{R}_T(f_L)$ auf Daten T.
- Starte Lernalgorithmus auf Daten D, gewinne so Hypothese f_D.
- Ausgabe: Hypothese f_D, benutze $\hat{R}_T(f_L)$ als Schätzer für das Risiko von f_D



Holdout-Testing: Analyse

- Ist der Schätzer $\hat{R}_T(f_L)$ für Risiko der Hypothese f_D
  - erwartungstreu,
  - optimistisch,
  - pessimistisch?

Holdout-Testing: Analyse

- Schätzer $\hat{R}_T(f_L)$ ist pessimistisch für $R(f_D)$:
  - $\hat{R}_T(f_L)$ ist erwartungstreu für f_L
  - f_L wurde mit weniger Trainingsdaten gelernt als f_D und hat daher im Erwartungswert ein höheres Risiko

- Aber Schätzer $\hat{R}_T(f_L)$ ist in der Praxis brauchbar, während $\hat{R}_L(f_L)$ meist 0 (oder nahe 0) ist.

Holdout-Testing: Analyse

- Warum wird Hypothese f_D trainiert und zurückgeliefert?
- Rückgegebene Hypothese f_D und nicht f_L, weil f_D ein geringeres Risiko hat, also besser ist.

Holdout-Testing: Analyse

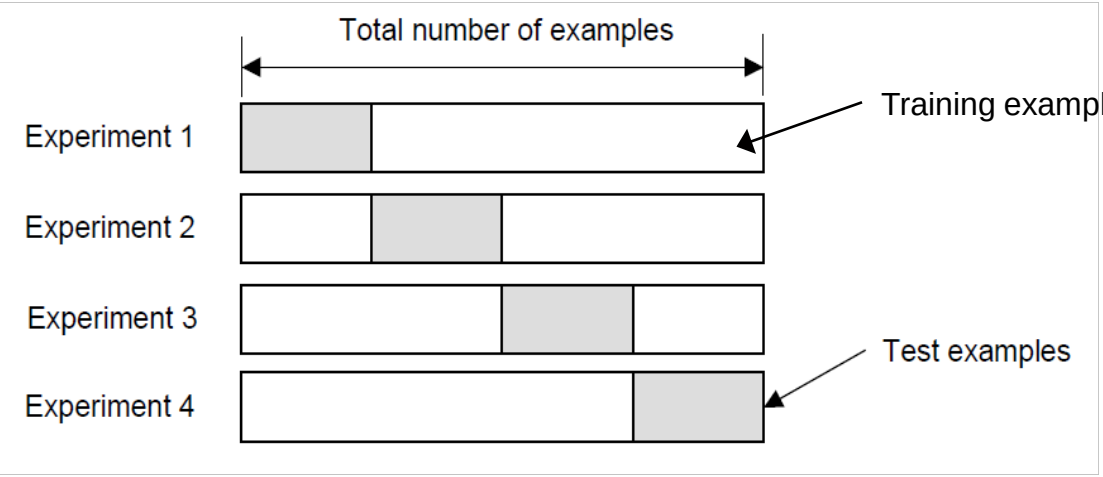
- Was sind die Vorteile/Nachteile wenn Testmenge T
  - möglichst groß
  - möglichst klein

  gewählt wird?
- Tmöglichst groß, damit Risikoschätzer $\hat{R}_T(f_L)$ geringe Varianz hat.
- Tmöglichst klein, damit Risikoschätzer $\hat{R}_T(f_L)$ geringen Bias hat, also nicht so stark pessimistisch ist.
- Braucht viele Daten, um gute Schätzungen zu bekommen
  - Wird praktisch nur verwendet, wenn sehr viele Daten zur Verfügung stehen.
  - Cross-Validation meist besser (siehe unten)

Cross-Validation

- Gegeben: Daten $D = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_d, y_d) \rangle$
- Teile D in n gleich große Abschnitte $D_1, ..., D_n$ mit $D = \bigcup_{i=1}^{n} D_i$ und $D_i \cap D_j = 0$
- Wiederhole für i=1..n
  - Trainiere f_i mit $L_i = D \setminus D_i$.
  - Bestimme empirisches Risiko $\hat{R}_{D_i}(f_i)$ auf D_i.

D_1	D_2	D_3	D_4

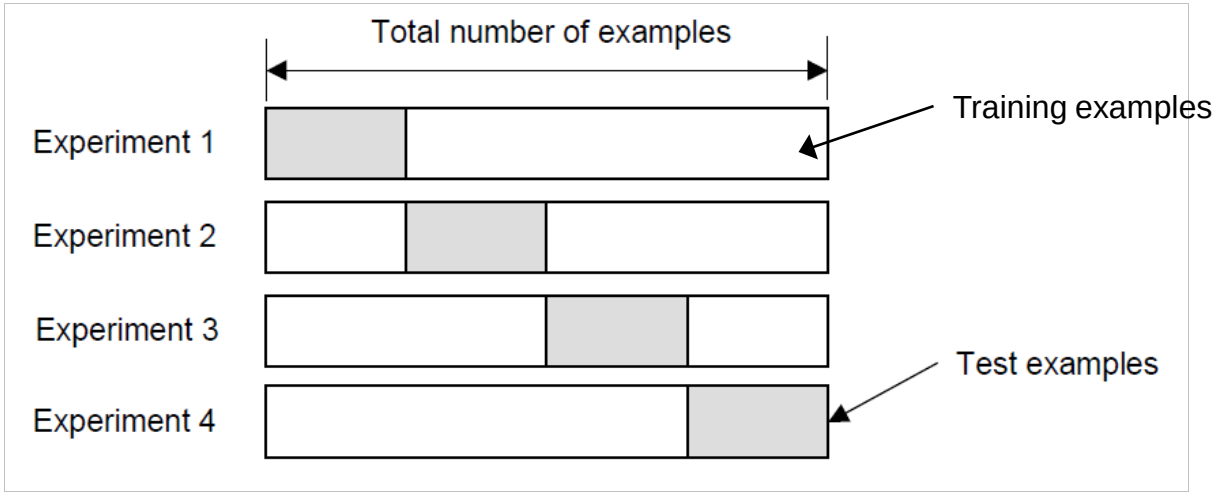


Cross-Validation

- Mittle empirische Risikoschätzungen auf den jeweiligen Testmengen D_i:

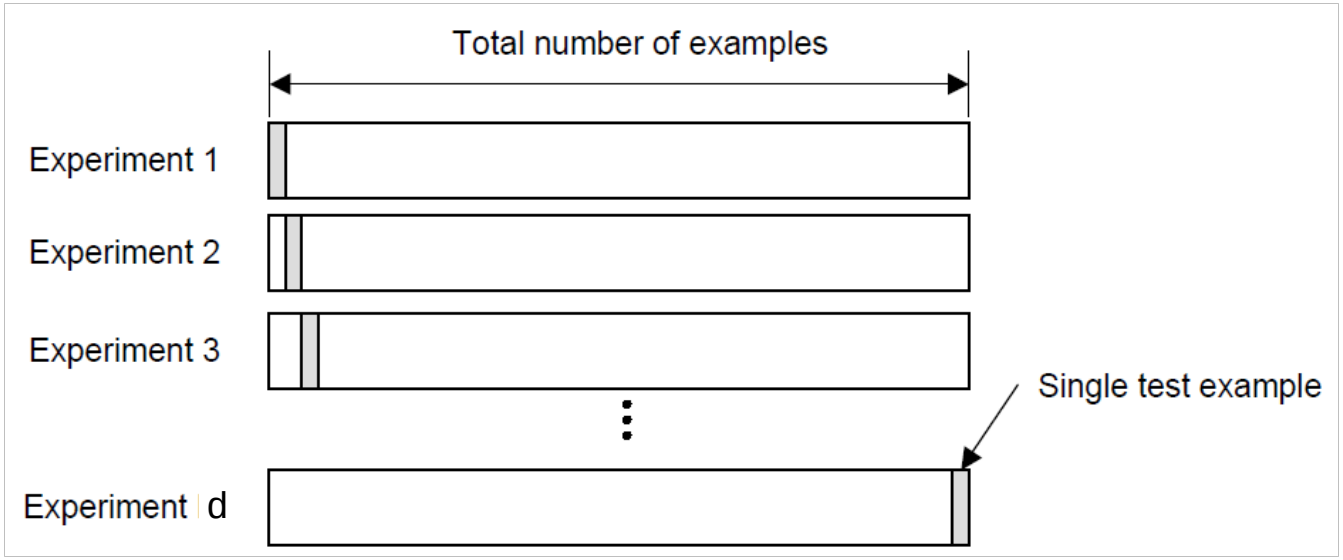
$$\overline{R} = \frac{1}{n} \sum_{i=1}^{n} \hat{R}_{D_i}(f_i)$$

- Trainiere f_D auf allen Daten D.
- Liefere Hypothese f_D und Schätzer $\overline{R}$.



Leave-One-Out Cross-Validation

- Spezialfall $n=d$ heisst auch *leave-one-out* Fehlerschätzung



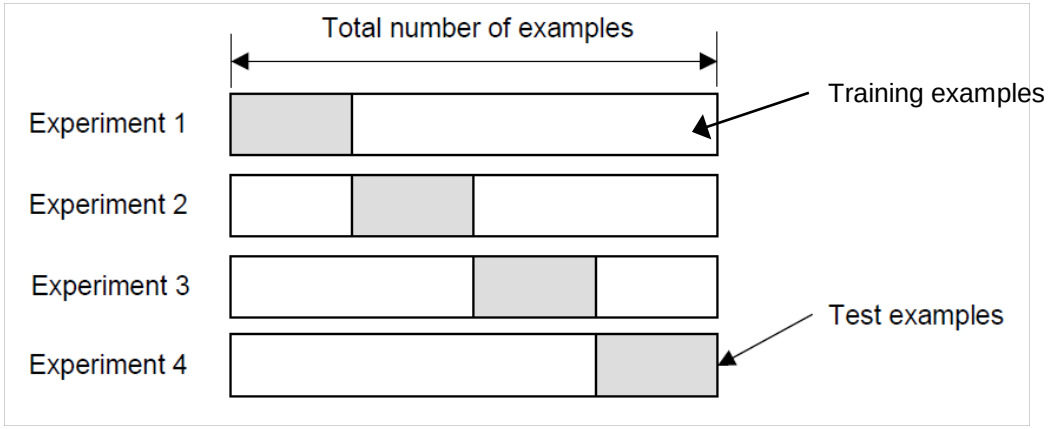
Cross-Validation: Analyse

- Ist der Schätzer
  - Optimistisch / pessimistisch / erwartungstreu?

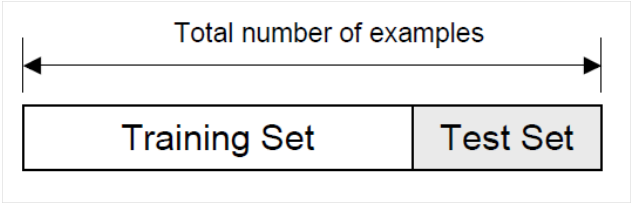
Cross-Validation: Analyse

- Ist der Schätzer
  - Optimistisch / pessimistisch / erwartungstreu?
- Schätzer ist pessimistisch:
  - Hypothesen f_i werden auf Anteil $(n\text{-}1)/n$ der verfügbaren Daten trainiert.
  - Hypothese f_D wird auf den gesamten Daten trainiert

Cross-Validation



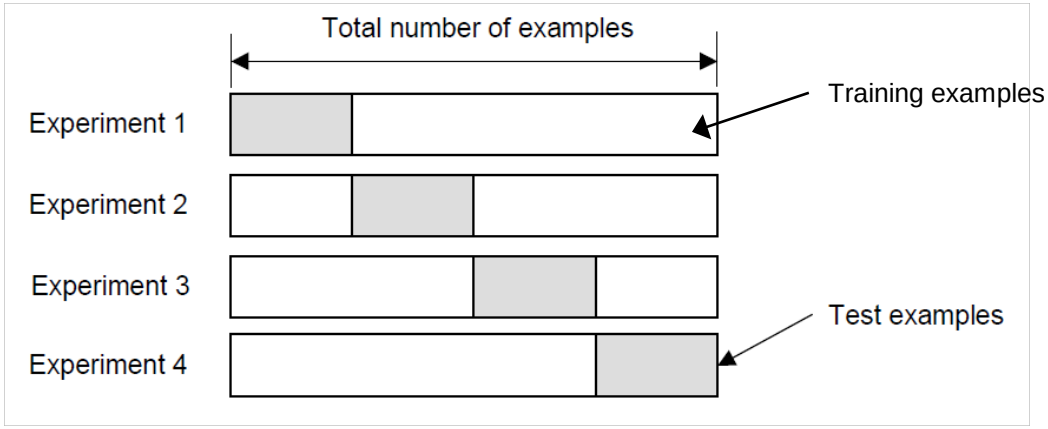
Holdout



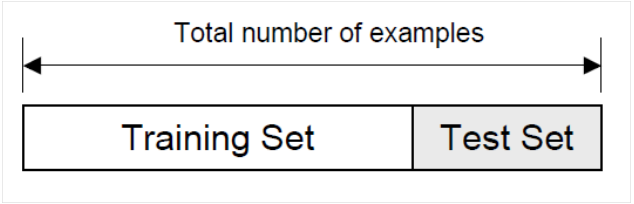
Cross-Validation: Analyse

- Bias/Varianz im Vergleich zu Holdout-Testing?
- Varianz ist geringer als beim Holdout-Testing
  - Mittelung über mehrere Holdout-Experimente, das reduziert die Varianz
  - Alle Daten fließen in den Schätzer ein
- Bias ähnlich wie beim Holdout-Testing, je nach Split-Verhältnissen.

Cross-Validation



Holdout



Überblick

- Hypothesenbewertung, Risikoschätzung
- Anwendungen/Beispiele
- Konfidenzintervalle
- Roc-Analyse

Anwendungen/Beispiele Hypothesenevaluierung

- Wir wollen/müssen aus Anwendungsgründen wissen, wie gut Hypothese ist
  - Macht es überhaupt Sinn, maschinelles Lernen in der Anwendung zu verwenden?
  - Mit wie vielen falschen Vorhersagen müssen wir rechnen?
- Wir wollen mehrere verschiedene Lernansätze vergleichen
  - Sollte man Entscheidungsbäume verwenden?
  - SVMs?
  - Logistische Regression?
  - Naive Bayes?

Anwendungen Hypothesenevaluierung

- Verfahren hat einen Parameter, den wir einstellen müssen
  - Regularisierungsparameter λ

$$f_{\mathbf{w}*} = \arg\min_{f_{\mathbf{w}}} \sum_i \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \lambda \| \mathbf{w} \|^2 \qquad \lambda=?$$

  - (Hyper)Parameter, der Modellklasse bestimmt, z.B. Polynomgrad bei polynomieller Regression

$$f_{\mathbf{w}}(x) = \sum_{i=0}^{M} w_i x^i \qquad \text{M=?}$$

- In allen diesen Fällen ist der Trainingsfehler kein geeignetes Entscheidungskriterium!
  - Fehlerschätzung mit Holdout-Menge oder Cross-Validation

Beispiel: Polynomgrad bei polynomieller Regression

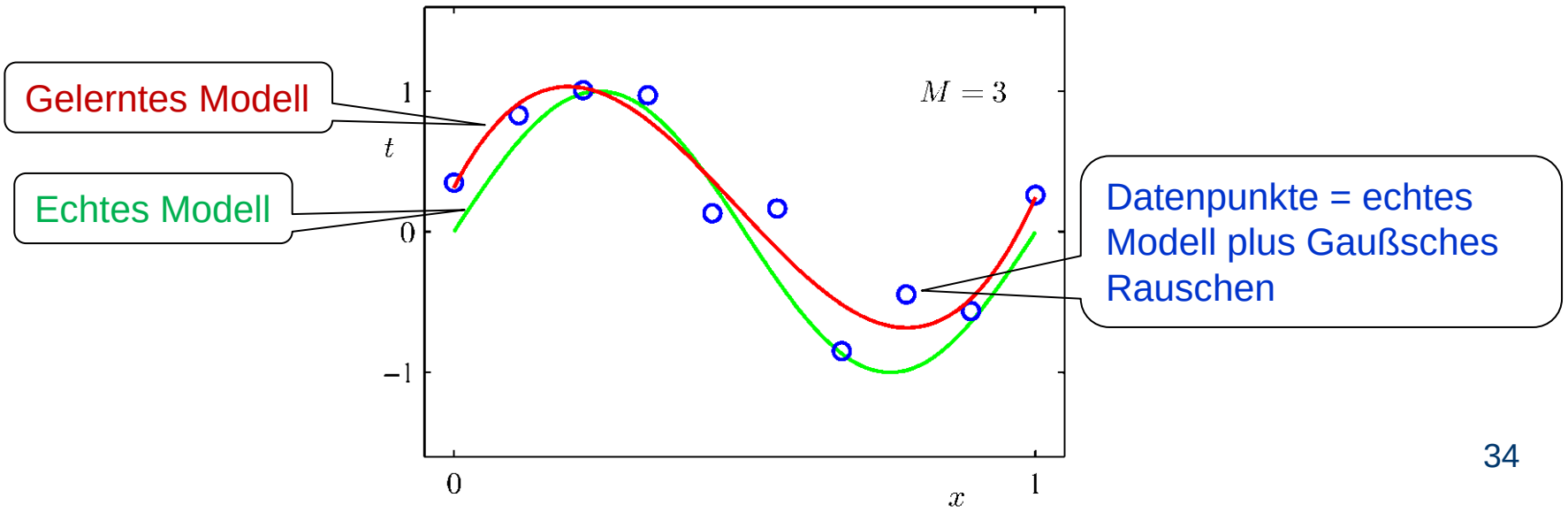
- Polynommodell vom Grad M: $f_{\mathbf{w}}(x) = \sum_{i=0}^{M} w_i x^i$

- Wir lernen Modell durch Minimierung des quadratischen Verlustes (unregularisiert – schlecht)

$$\mathbf{w}_* = \arg\min_{\mathbf{w}} \sum_{i=1}^{m} (f_{\mathbf{w}}(x_i) - y_i)^2$$

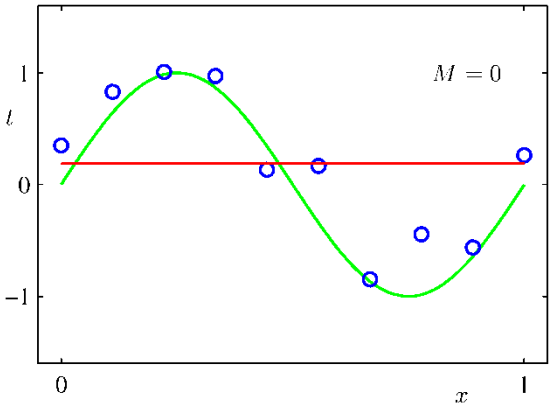
Trainingsdaten

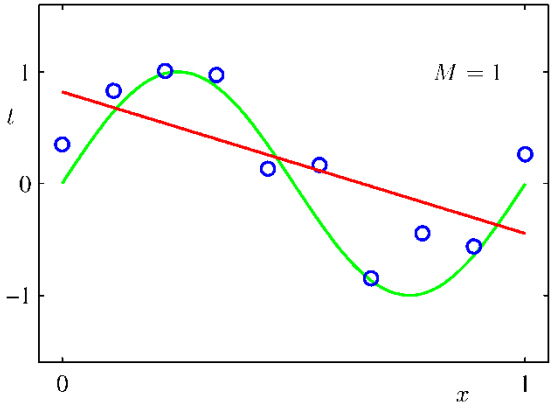
$$L = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$$

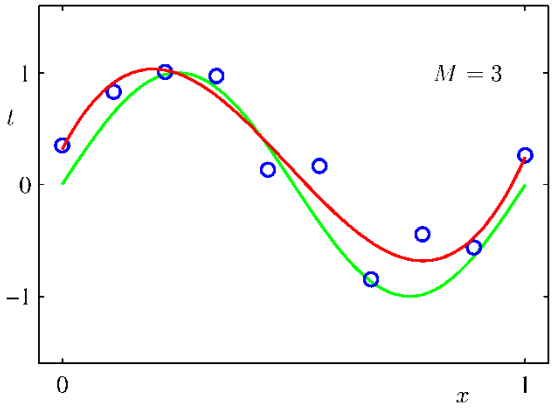


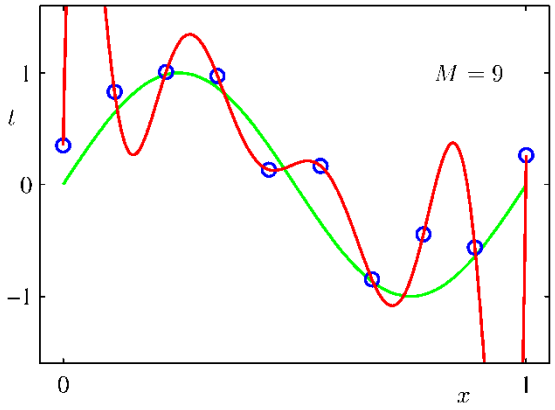
Beispiel polynomielle Regression: Training vs. Testfehler

- Erfolg des Lernens hängt vom gewählten Polynomgrad M ab, der Komplexität des Modells kontrolliert



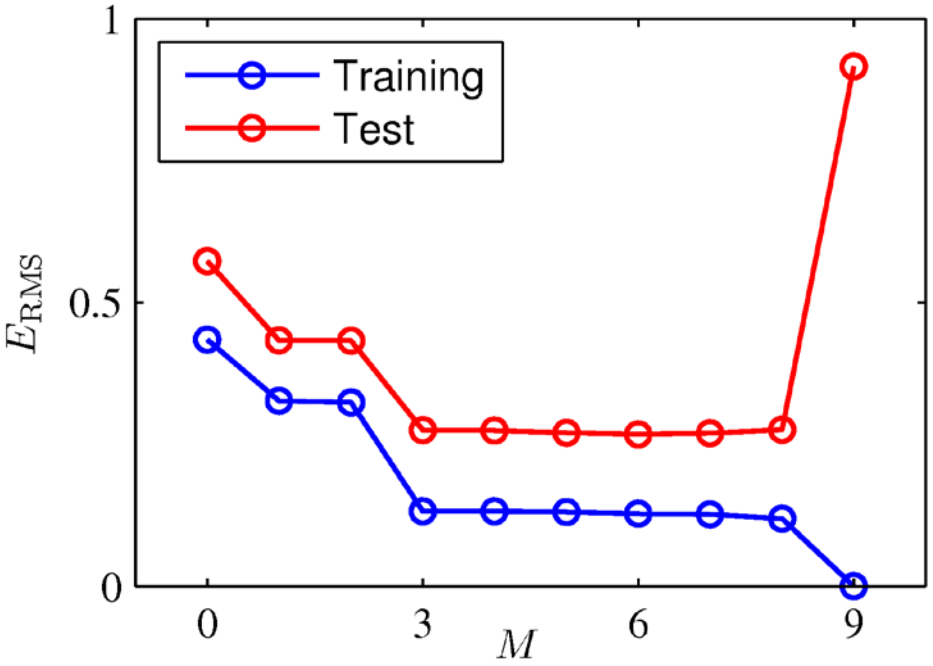






Beispiel polynomielle Regression: Training vs. Testfehler

- Trainingsfehler vs. Testfehler für unterschiedliche Polynomgrade
  - Testfehler z.B. mit grosser Holdout-Menge gemessen

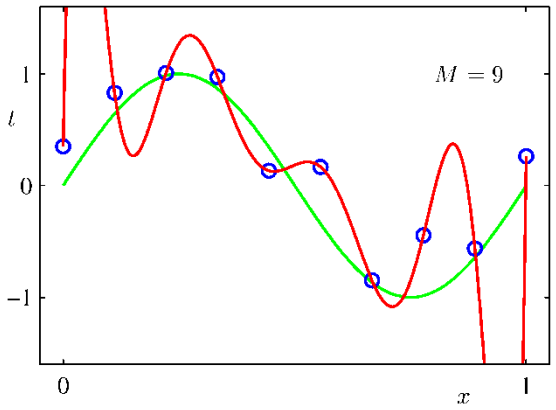


Phänomen der Überanpassung („Overfitting“):
Trainingsfehler sinkt monoton mit M, aber Testfehler steigt für große M wieder

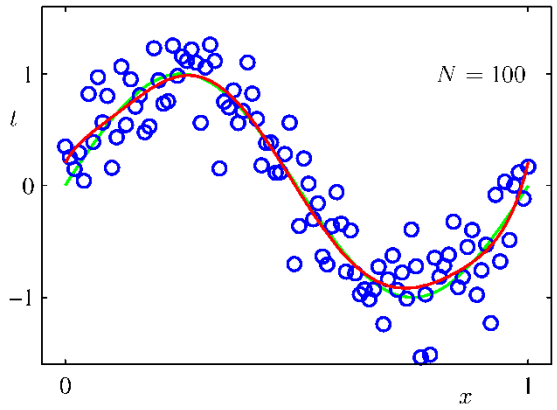
Beispiel polynomielle Regression: Training vs. Testfehler

- Problem der Überanpassung (Overfitting) wird geringer, wenn mehr Daten verfügbar

10 Beispiele, M = 9



100 Beispiele, M = 9



- In der Praxis verfügbare Daten begrenzt
  - Modell muss regularisiert werden

Regularisierte Polynomielle Regression

- Polynommodell vom Grad $M=9$: $f(x) = \sum_{i=0}^{M} w_i x^i$

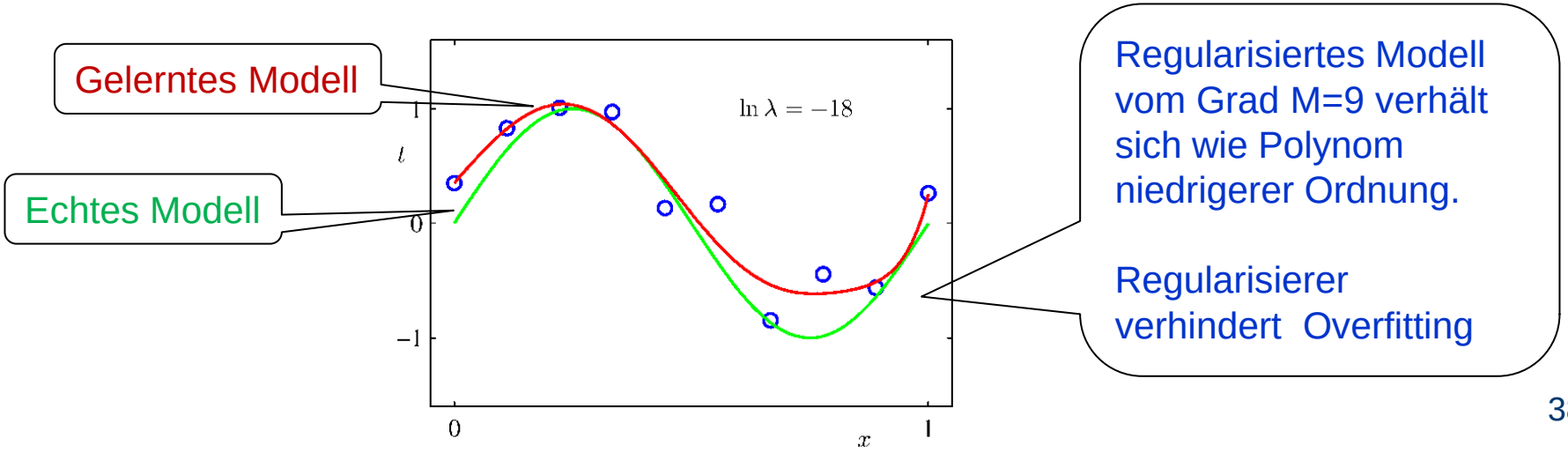
- Lerne Modell durch Minimierung des regularisierten quadratischen Verlustes

$$\mathbf{w}_* = \arg\min_{\mathbf{w}} \sum_{i=1}^{M} (f_{\mathbf{w}}(x_i) - y_i)^2 + \lambda \|\mathbf{w}\|^2$$

Trainingsdaten

$$L = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$$

$$M = 9, \quad \ln \lambda = -18$$



Regularisiertes Modell vom Grad M=9 verhält sich wie Polynom niedrigerer Ordnung.

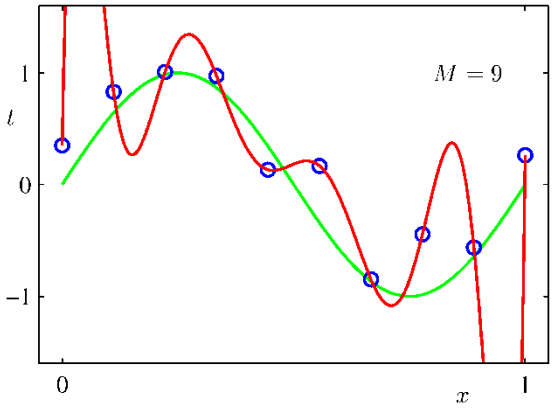
Regularisierer verhindert Overfitting

Regularisierte Polynomielle Regression

- Jetzt müssen wir λ einstellen
- Regularisierungsparameter kontrolliert Komplexität ähnlich wie Polynomgrad M

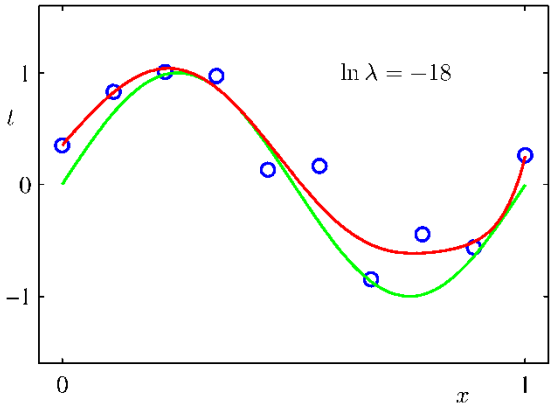
$M = 9$

$\ln \lambda = -\infty$



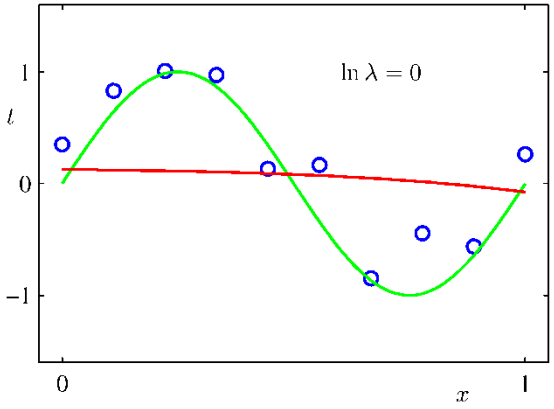
$M = 9$

$\ln \lambda = -18$



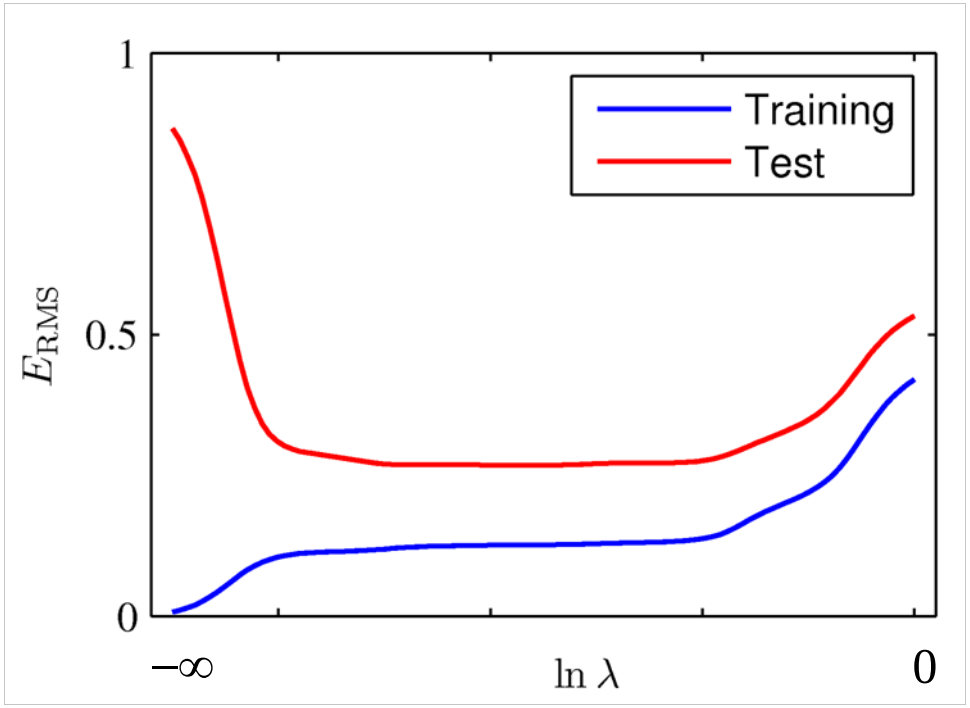
$M = 9$

$\ln \lambda = 0$



Regularisierte Polynomielle Regression

- Trainingsfehler vs. Testfehler für unterschiedliche Werte des Regularisierungsparameters λ
  - Testfehler z.B. mit grosser Holdout-Menge gemessen

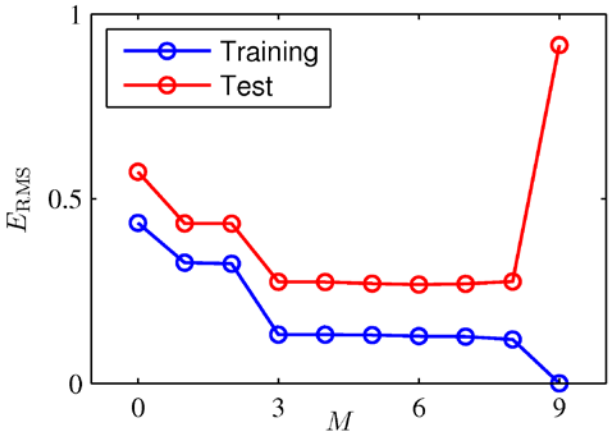


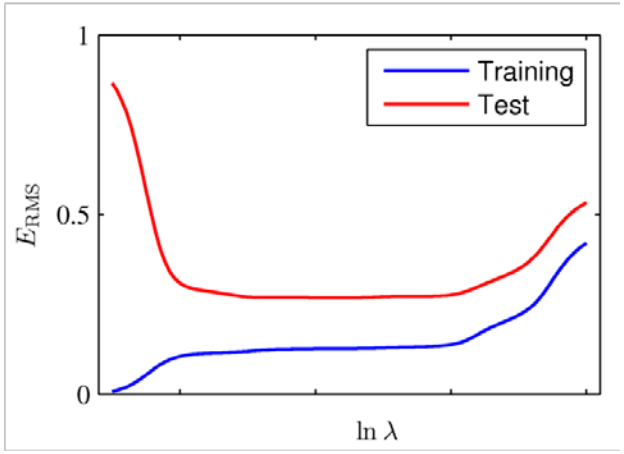
Trainingsfehler minimal ohne Regularisierung, aber Testfehler besser für regularisiertes Modell

Umgekehrtes Verhalten wie für unterschiedliche Polynomgrade

Regularisierte Polynomielle Regression

- Regularisierer wirkt wie eine Begrenzung der Modellkomplexität und verhindert Überanpassung
- In der Praxis am besten, Modellkomplexität durch Regularisierung zu kontrollieren (direkter Parameter wie bei Polynomen oft nicht verfügbar)
- Regularisierer kann durch Fehlerschätzung (Holdout-Testing oder Cross-Validation) eingestellt werden.





Triple-Cross-Validation

- Ziel: Abschätzung der Genauigkeit von Vorhersagen unter optimalen Parametern

Triple-Cross-Validation

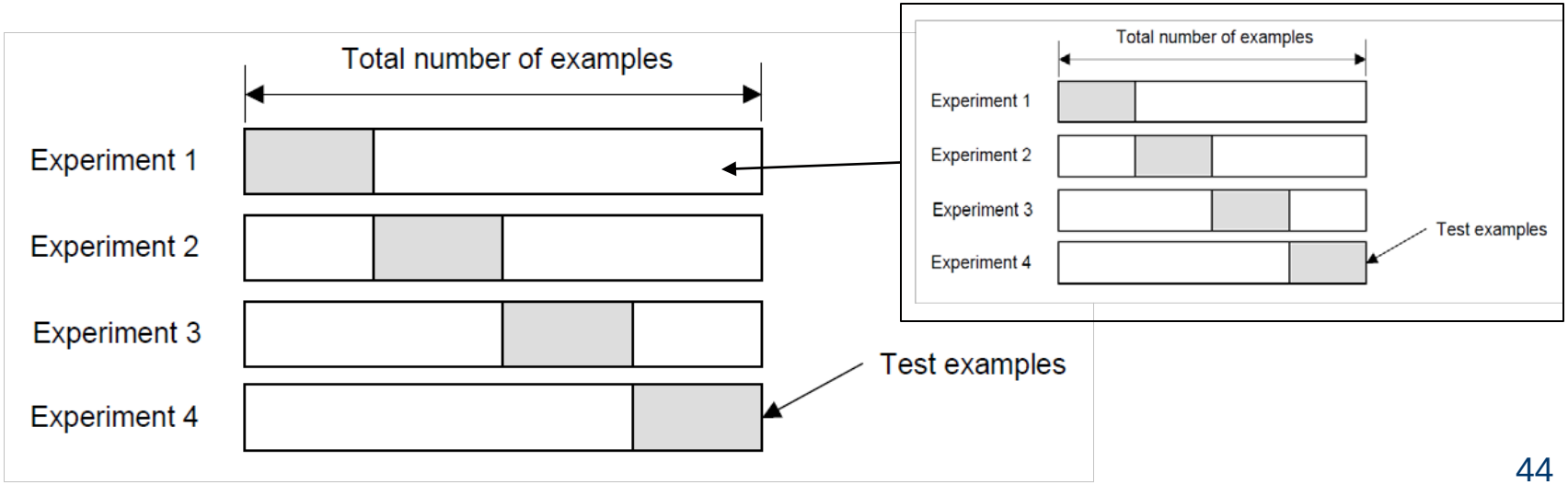
- Gegeben: Daten $D = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_d, y_d) \rangle$
- Teile D in n Abschnitte $D_i = \langle (\mathbf{x}_{i_1}, y_{i_1}), ..., (\mathbf{x}_{i_k}, y_{i_k}) \rangle, \quad k = d / n$ mit $D = \bigcup_{i=1}^{n} D_i$ und $D_i \cap D_j = 0$
- Wiederhole für i=1..n
  - Teile $D \setminus D_i$ in m Abschnitte mit $D \setminus D_i = \bigcup_{j=1}^{m} D_{i,j}$ und $D_{i,j} \cap D_{i,k} = 0$
  - Wiederhole für j=1..m
    - Trainiere $f_{i,j,C}$ mit $D \setminus D_i \setminus D_{i,j}$ f.a. möglichen Parameter C
    - Bestimme empirisches Risiko $\hat{R}_C(f_{i,j,C})$ auf $D_{i,j}$.
    - Bestimme C^* mit minimalen Risiko $\hat{R}_C$
  - Trainiere f_i mit $D \setminus D_i$ und C^*
  - Bestimme empirisches Risiko $\hat{R}_{D_i}(f_i)$ auf D_i.

Triple-Cross-Validation

- Mittle empirische Risikoschätzungen auf den jeweiligen Testmengen D_i:

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} \hat{R}_{D_i}(f_i)$$

- Trainiere f_D auf allen Daten D.
- Liefere Hypothese f_D und Schätzer $\overline{R}$.



Überblick

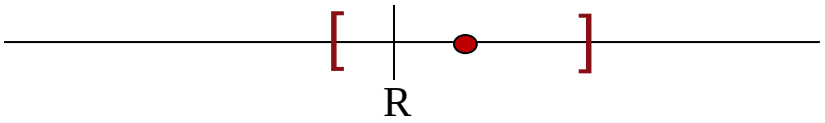
- Hypothesenbewertung, Risikoschätzung
- Anwendungen/Beispiele
- Konfidenzintervalle
- Roc-Analyse

Konfidenzintervalle

- Idee Konfidenzintervall:
  - Intervall um den geschätzten Fehler $\hat{R}(f)$ angeben
  - so dass der echte Fehler „meistens“ im Intervall liegt
  - Quantifiziert Unsicherheit der Schätzung

- Weg zum Konfidenzintervall: Analyse der Verteilung der Zufallsvariable $\hat{R}(f)$

$\hat{R}(f)$



Zero-One Loss und Fehlerwahrscheinlichkeit

- Für Konfidenzintervalle betrachten wir Risikoschätzung im Spezialfall Klassifikation mit Zero-One Loss
- Verlustfunktion Zero-One Loss:
  - $\ell(y, y') = \begin{cases} 0, & wenn \quad y = y' \\ 1, & wenn \quad y \neq y' \end{cases}$
- $\rightarrow$ Risiko = Fehlerwahrscheinlichkeit.
  - $R(f) = \int \ell(y, f(\mathbf{x}))\, p(\mathbf{x}, y) d\mathbf{x} dy$

    $= \int [[y \neq f(\mathbf{x})]]\, p(\mathbf{x}, y) d\mathbf{x} dy$

    $= p(y \neq f(\mathbf{x}))$

    $[[Ereignis]]$: binäre Indikatorvariable für "Ereignis"

Verteilung für Fehlerschätzer

- Hypothese f wird auf separater Testmenge mit m unabhängigen Beispielen evaluiert:

$$\hat{R}_T(f) = \frac{1}{m}\sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j)) \qquad T = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m) \rangle$$

- Fehlerschätzer ist erwartungstreu, $E[\hat{R}_T(f)] = R(f)$

- Betrachten zunächst unnormalisierten Fehlerschätzer

$$m \cdot \hat{R}_T(f) = \sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j))$$

- Summe über Beispielverluste $\ell(y_j, f(\mathbf{x}_j)) \in \{0,1\}$
- Beispiele unabhängig: Summe über Münzwürfe
- Münzparameter ist Fehlerwahrscheinlichkeit $R(f)$

Binomialverteilung

- Unnormalisiertes empirisches Risiko

$$m \cdot \hat{R}_T(f) = \sum_{j=1}^{m} \ell(y_j, f(\mathbf{x}_j))$$

ist Summe von Bernouilli-Variablen, also binomialverteilt:

$$m\hat{R}_T(f) \sim \text{Bin}\left(m\hat{R}_T(f) \mid m, r\right) \qquad r = R(f)$$

- Erwartungswert $E\left[m \cdot \hat{R}_T(f)\right] = mr$
- Varianz $Var\left[m \cdot \hat{R}_T(f)\right] = m \cdot r(1-r)$

Normalisierte Binomialverteilung

- Normalisierter Fehlerschätzer: normalisierte Binomialverteilung

- Erwartungswert des normalisierten Fehlerschätzers:

$$E\left[\hat{R}_T(f)\right] = \frac{1}{m} E\left[m\hat{R}_T(f)\right] = \frac{1}{m} mr = r$$

- Varianz des normalisierten Fehlerschätzers:

$$Var\left[\hat{R}_T(f)\right] = \frac{1}{m^2} Var\left[m\hat{R}_T(f)\right] = \frac{1}{m^2} m \cdot r(1-r) = \frac{r(1-r)}{m}$$

- Standardabweichung („Standardfehler“)

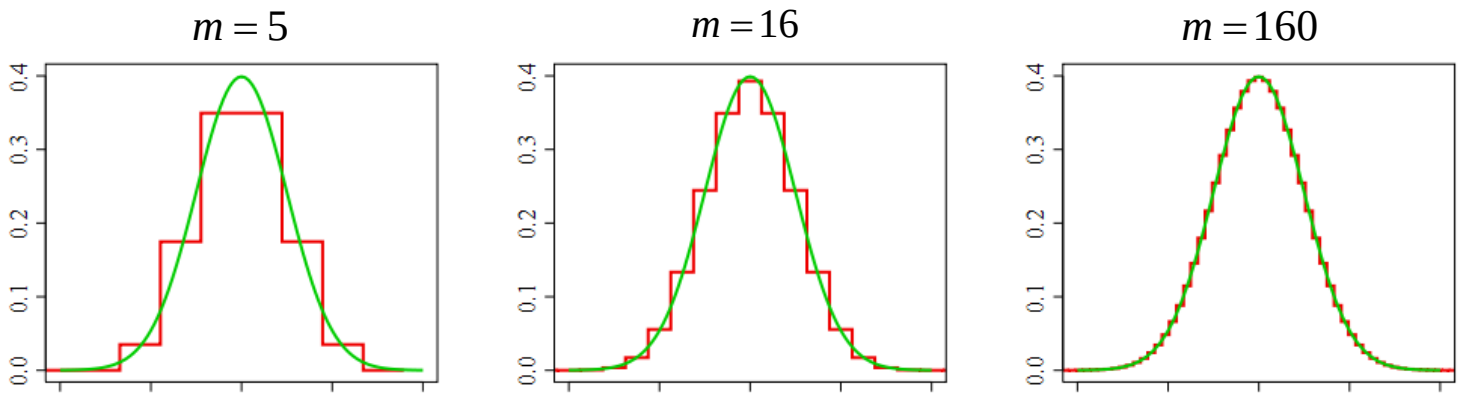
$$\sigma_{\hat{r}} = \sqrt{\frac{r(1-r)}{m}}$$

  - Zufallsanteil des Schätzers, sinkt mit $1/\sqrt{m}$

Binomialverteilung

- Binomialverteilung für große m ähnlich Normalverteilung

$r = 0.5$



Normalverteilung

- Empirisches Risiko annähernd normalverteilt:

$$\hat{R}_T(f) \sim N\left(\hat{R}_T(f) \mid r, \sigma_{\hat{r}}^2\right) \quad \text{[approximativ, für große m]}$$

$$\sigma_{\hat{r}}^2 = \frac{r(1-r)}{m}$$

- Für die weitere Analyse betrachten wir das standardisierte Risiko, dieses ist standardnormalverteilt:

$$\frac{\hat{R}_T(f) - R(f)}{\sigma_{\hat{r}}} \sim N\left(\frac{\hat{R}_T(f) - R(f)}{\sigma_{\hat{r}}} \mid 0,1\right) \quad \text{[approximativ, für große m]}$$

- Schätzen der Varianz des empirischen Risikos:

$$\sigma_{\hat{r}}^2 \approx s_{\hat{r}}^2 \qquad s_{\hat{r}}^2 = \frac{\hat{r}(1-\hat{r})}{m-1}, \qquad \hat{r} = \hat{R}_T(f)$$

Schranken für echtes Risiko

- Wir haben Verteilung des empirischen Risikos hergeleitet
- Was sagt das empirische Risiko jetzt also über das echte Risiko?
- Wir können Schranke für echtes Risiko bestimmen!

$$P\left(R(f) \leq \hat{R}_T(f) + \varepsilon\right) = P\left(R(f) - \hat{R}_T(f) \leq \varepsilon\right) = P\left(\frac{R(f) - \hat{R}_T(f)}{s_{\hat{r}}} \leq \frac{\varepsilon}{s_{\hat{r}}}\right) \approx \Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right)$$

Approximativ standardnormalverteilt

$\Phi(z)$ kumulative Verteilungsfunktion der Standardnormalverteilung

Kumulative Verteilungsfunktion der Normalverteilung

- Sei Z standardnormalverteilte Zufallsvariable

$$Z \sim \mathcal{N}(Z \mid 0,1)$$

- Kumulative Verteilungsfunktion der Normalverteilung:

$$\begin{aligned}\Phi(z) &= p(Z \leq z) \\ &= \int_{-\infty}^{z} \mathcal{N}(x \mid 0,1)dx \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)dx\end{aligned}$$

- Keine geschlossene Lösung, nachschlagen in Tabelle

Schranken für echtes Risiko

- Ergebnis: Einseitige Schranke für echtes Risiko, gegeben ε

$$P\left(R(f) \le \hat{R}_T(f) + \varepsilon\right) = P\left(R(f) - \hat{R}_T(f) \le \varepsilon\right) \approx \Phi\left(\frac{\varepsilon}{s_{\hat{f}}}\right)$$

Kann man in der Praxis bestimmen:
$s_{\hat{f}}$ beobachtet

- Zweiseitige Schranke:

$$P\left(|R(f) - \hat{R}_T(f)| \le \varepsilon\right) = \ldots$$

$$= \ldots$$

$$\approx 2\Phi\left(\frac{\varepsilon}{s_{\hat{f}}}\right) - 1$$

Konfidenzintervalle

- Idee Konfidenzintervall: ε so wählen, dass Schranke mit vorgegebener Wahrscheinlichkeit von 1-δ (z.B. δ =0.05) gilt.

- Einseitiges 1-δ-Konfidenzintervall: Schranke ε, so dass

$$P\left(R(f) \leq \hat{R}_T(f) + \varepsilon\right) \geq 1 - \delta$$

- Zweiseitiges 1-δ-Konfidenzintervall: Schranke ε, so dass

$$P\left(|R(f) - \hat{R}_T(f)| \leq \varepsilon\right) \geq 1 - \delta$$

- Bei symmetrischer Verteilung gilt immer:
  - ε zu einseitigem 1-δ-Konfidenzintervall
    = ε zu zweiseitigem 1-2δ-Konfidenzintervall.

5% Wahrscheinlichkeit, dass $R(f) > \hat{R}_T(f) + \varepsilon$

10% Wahrscheinlichkeit, dass $R(f) > \hat{R}_T(f) + \varepsilon$ oder $R(f) < \hat{R}_T(f) - \varepsilon$

Konfidenzintervalle

- Ermitteln des einseitigen 1-δ-Konfidenzintervalls:

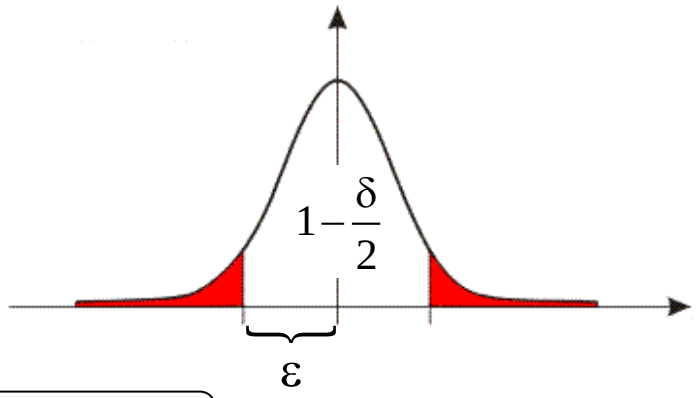
$$P\left(R(f) \le \hat{R}_T(f) + \varepsilon\right) \ge 1 - \delta$$

$$\tilde{\Leftrightarrow} \; \Phi\left(\frac{\varepsilon}{s_{\hat{r}}}\right) \ge 1 - \delta$$

$$\Leftrightarrow \frac{\varepsilon}{s_{\hat{r}}} \ge \Phi^{-1}(1-\delta)$$

$$\Leftrightarrow \varepsilon \ge s_{\hat{r}} \Phi^{-1}(1-\delta)$$

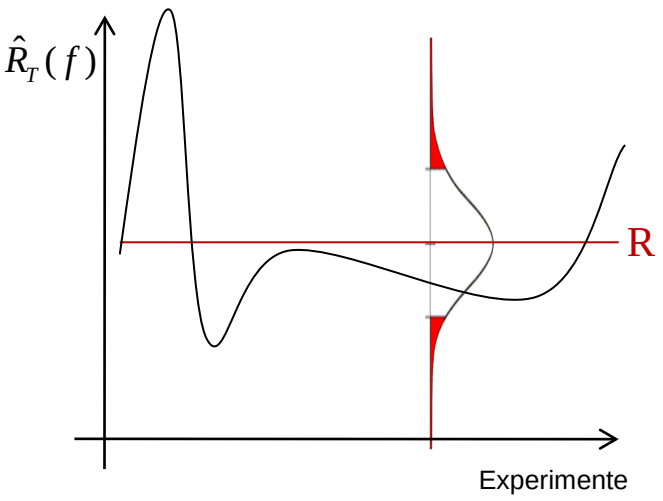
$\Phi(z)$ streng monoton

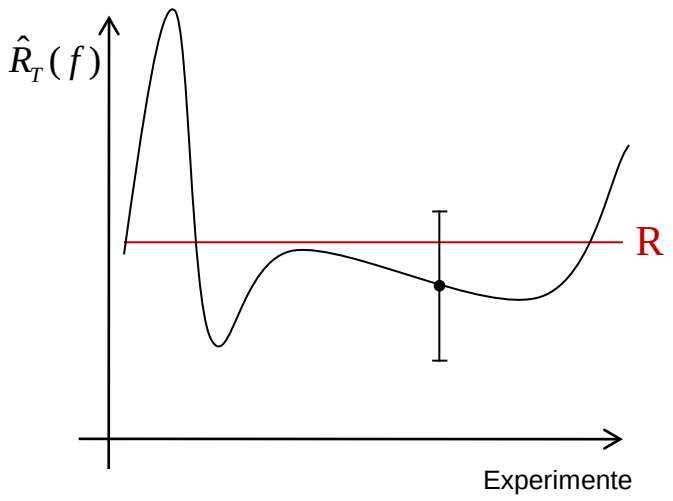


- Konfidenzintervall ist $[\hat{R}_T(f) - \varepsilon, \hat{R}_T(f) + \varepsilon]$ (Konfidenzlevel 1-2δ)

Konfidenzintervalle

- $\hat{R}_T(f)$ ist annähernd normal-verteilt





Konfidenzintervalle

- Beispiel:

$\hat{r} = 0.08$ beobachtetes empirisches Risiko, m=100

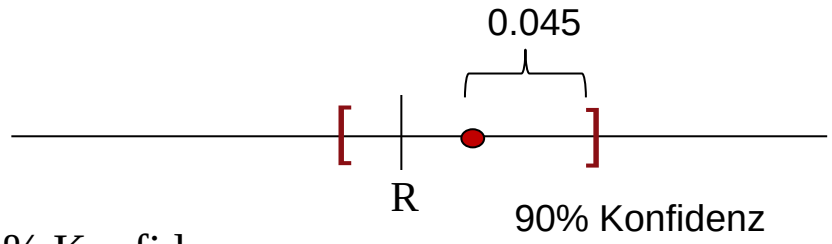
$s_{\hat{r}} = \sqrt{\frac{0.08 \cdot 0.92}{100 - 1}} \approx 0.027$ empirischer Standardfehler

$\delta = 0.05$ Konfidenzlevel

$\varepsilon \geq s_{\hat{r}} \Phi^{-1}(1-\delta)$

$\varepsilon \geq 0.027 \cdot 1.645 \approx 0.045$

$\Phi^{-1}(0.95)$ [Tabelle]



Einseitige Schranke: $R(f) \leq R_T(f) + \varepsilon$ mit 95% Konfidenz

Zweiseitige Schranke: $\left|R(f) \leq R_T(f)\right| \leq \varepsilon$ mit 90% Konfidenz

Students t-Verteilung

- Empirisches Risiko annähernd normalverteilt:
  - $$p\left(\hat{R}_T(f)=\hat{r}\mid r\right)=B\left(m\hat{r}\mid r,m\right)$$
    $$\approx N\left(\hat{r}\mid r,\frac{r(1-r)}{m}\right)$$
    $$=N\left(\frac{\hat{r}-r}{\sigma_{\hat{r}}}\mid 0,1\right)$$

    Einfache Charakterisierung der Verteilung des empirischen Fehlers
- Problem: Risiko muss bekannt sein, damit wir Varianz bzw. Standardfehler bestimmen können.
  - $$\sigma_{\hat{r}}^2=\frac{r(1-r)}{m};\quad \sigma_{\hat{r}}=\sqrt{\frac{r(1-r)}{m}}$$
- Nur das empirische Risiko ist gegeben.

Students t-Verteilung

- Standardisiertes empirisches Risiko mit geschätzter Varianz

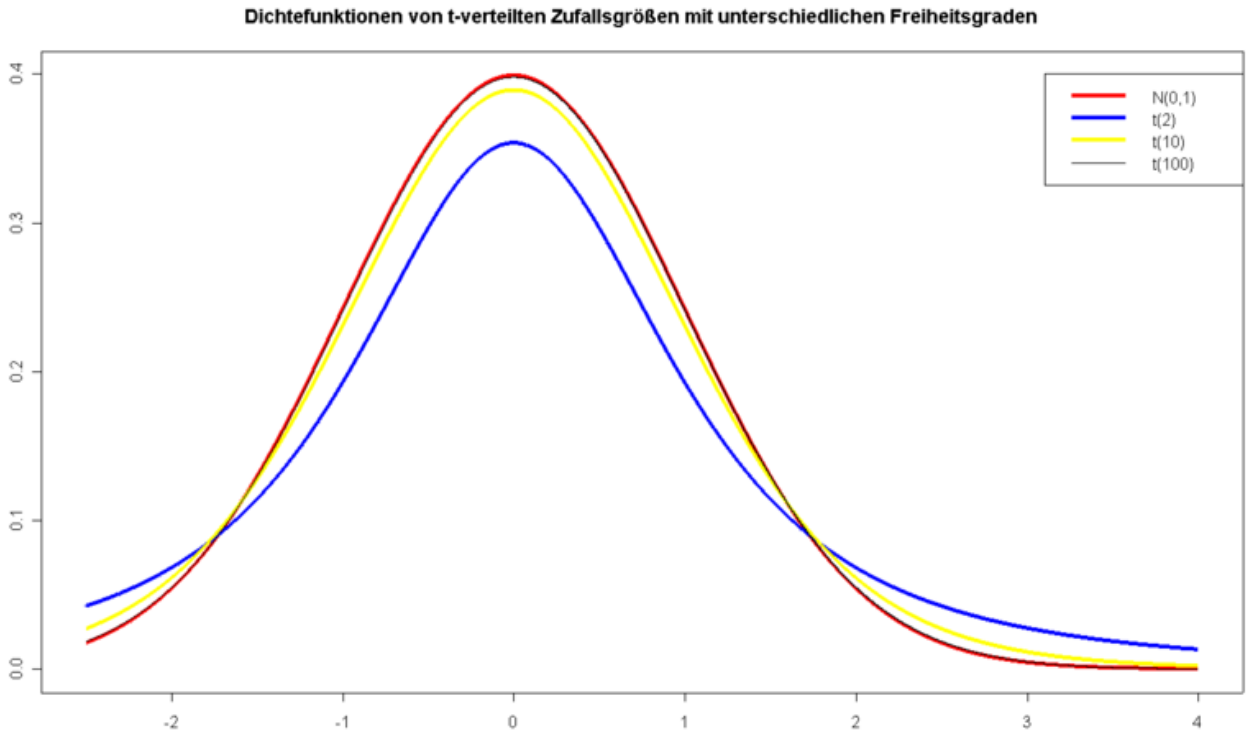
$$\frac{\hat{R}_T(f) - R(f)}{s_{\hat{f}}}$$

folgt einer Students t-Verteilung mit $m\text{-}1$ Freiheitsgraden (ähnlich Normalverteilung, aber mehr Wahrscheinlichkeitsmasse in den Außenbereichen).

- Aber für große m konvergiert Students t-Verteilung wieder gegen die Standardnormalverteilung

$$\frac{\hat{R}_T(f) - R(f)}{s_{\hat{f}}} \sim N\left(\frac{\hat{R}_T(f) - R(f)}{s_{\hat{f}}} \mid 0,1\right) \quad \text{[approximativ, für große m]}$$

Students t-Verteilung



$$\lim_{m\to\infty} t\left(\frac{\hat{r}-r}{s_{\hat{r}}}|m\right) = N\left(\frac{\hat{r}-r}{s_{\hat{r}}}|0,1\right)$$

$$f_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t$$

Konfidenzintervalle

- Vorsicht bei der Interpretation von Konfidenzintervallen: die Zufallsvariable ist das empirische Risiko $\hat{r}$ und das davon abgeleitete Intervall ε, nicht das echte Risiko $R(f)$.

- Richtig:

"Die Wahrscheinlichkeit, bei einem Experiment ein Konfidenzintervall zu erhalten, das den echten Fehler enthält, ist 90%"

- Falsch:

"Wir haben ein Konfidenzintervall ε erhalten. Die Wahrscheinlichkeit, dass der echte Fehler im Intervall liegt, ist 90%"

Überblick

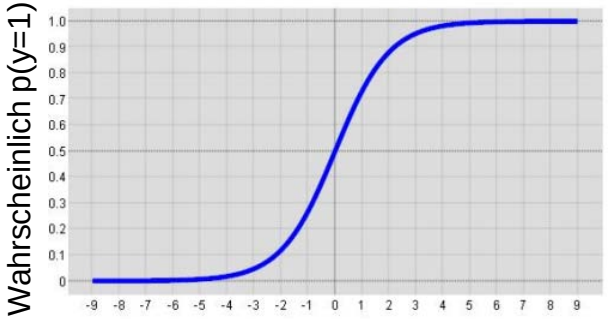
- Hypothesenbewertung, Risikoschätzung
- Anwendungen/Beispiele
- Konfidenzintervalle
- Roc-Analyse

Klassifikator / Entscheidungsfunktion

- Für eine binäre Klassifikation (y= +1 oder -1) wird oft eine kontinuierliche Entscheidungsfunktion $f(x)$ gelernt.
  - Z.B. lineares Modell

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{m} w_i x_i$$

- Je größer $f(x)$, desto wahrscheinlicher ist, dass x zur Klasse +1 gehört
  - Z.B. logistische Regression



$$\sigma(f(x)) = \frac{1}{1+\exp(-f(x))}$$

Klassifikator / Entscheidungsfunktion

- Wie bestimmen wir Klassenentscheidung +1/-1 aus f(x)?
- Allgemeine Lösung:

$$\text{Vorhersage} = \begin{cases} +1 : f(\mathbf{x}) \geq \theta \\ -1 : \text{sonst} \end{cases}$$

- Der Wert für θ verschiebt „false positives“ zu „false negatives“.
- Optimaler Wert hängt von Kosten einer positiven oder negativen Fehlklassifikation ab.

Evaluation von Klassifikatoren und Entscheidungsfunktionen

- Fehlklassifikationswahrscheinlichkeit
  - Häufig nicht aussagekräftig, weil P(+1) sehr klein.
  - Wie gut sind 5% Fehler, wenn P(+1)=3%?
  - Idee: Nicht Klassifikator bewerten, sondern Entscheidungsfunktion.

- Receiver Operating Characteristic (ROC-Kurve)
  - Bewertet Entscheidungsfunktion,
  - Jeder Punkt auf der ROC Kurve entspricht einem Schwellwert θ
  - Fläche unter ROC-Kurve = P(positives Beispiel hat höheren f-Wert als negatives Beispiel)

ROC-Analyse

- Entscheidungsfunktion + Schwellwert = Klassifikator

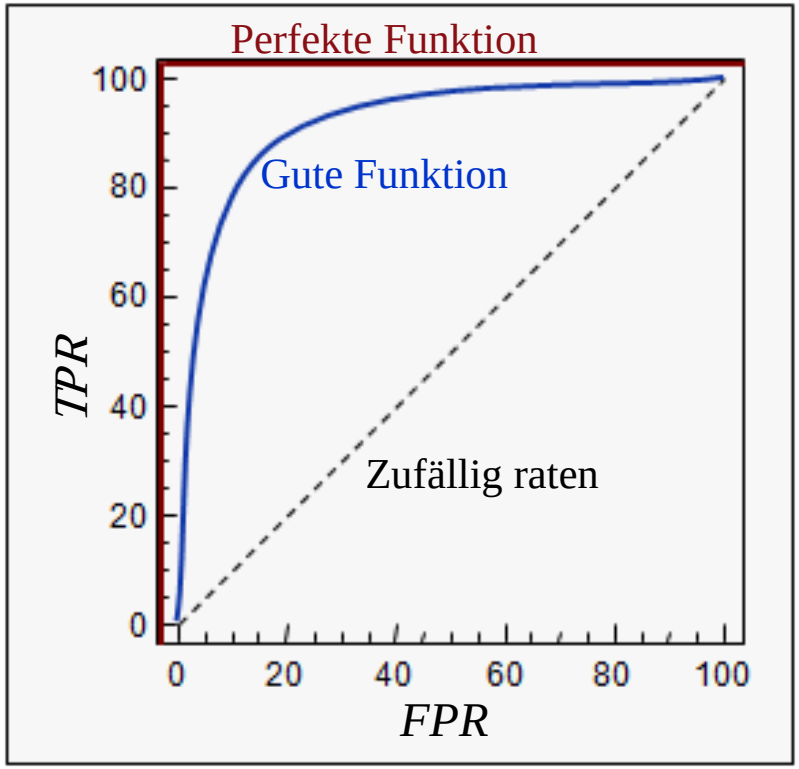
$$\text{Vorhersage} = \begin{cases} +1 : f(\mathbf{x}) \geq \theta \\ -1 : \text{sonst} \end{cases}$$

  - Fehler hängen vom Schwellwert ab
  - Großer Schwellwert: Mehr positive Bsp falsch.
  - Kleiner Schwellwert: Mehr negative Bsp falsch.

- ROC-Analyse: Bewertung der Entscheidungsfunktion unabhängig vom konkreten Schwellwert.
- Charakterisieren das Verhalten des Klassifikators für alle möglichen Schwellwerte.

ROC-Kurven

- Rate der „False Positives“ und „True Positives“ in Abhängigkeit des Schwellwertes
  - X-Achse: „False Positive Rate“
  - Y-Achse: „True Positive Rate“



	Vorhersage „+“	Vorhersage „-“
Echtes Label „+“	TP	FN
Echtes Label „-“	FP	TN

$$FPR = \frac{FP}{N} \qquad TPR = \frac{TP}{P}$$

$$N = FP + TN \qquad P = TP + FN$$

Bestimmen der ROC-Kurve von f

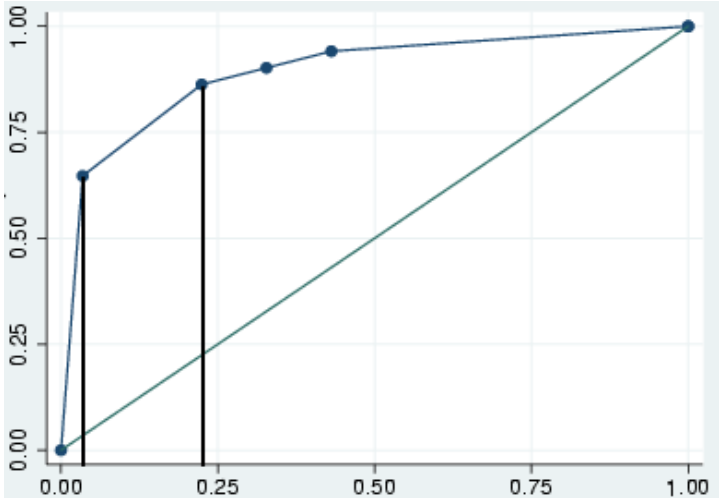
- Annahme: kein $f(\boldsymbol{x}) = f(\boldsymbol{x}')$ für $\boldsymbol{x} \neq \boldsymbol{x}'$.

- Generiere Liste L aller Instanzen $\boldsymbol{x}$, absteigend sortiert nach $f(\mathbf{x})$
- P = Anzahl positiver Instanzen, N = Anzahl negativer Instanzen
- TP = FP = 0

- Für i = 1 bis Länge(L)
  - x = i-tes Element von L
  - Wenn x positive Instanz: increment(TP)
  - Wenn x negative Instanz: increment(FP)
  - Zeichne neuen Punkt mit Koordination (FP/N,TP/P)

Flächeninhalt der ROC-Kurve

- Flächeninhalt AUC kann durch Integrieren (Summieren der Trapez-Flächeninhalte) bestimmt werden.



- x_+ = zufällig gezogenes Positivbeispiel
- x_- = zufällig gezogenes Negativbeispiel
- Theorem: AUC = $P(f(x_+) > f(x_-))$.

Precision / Recall

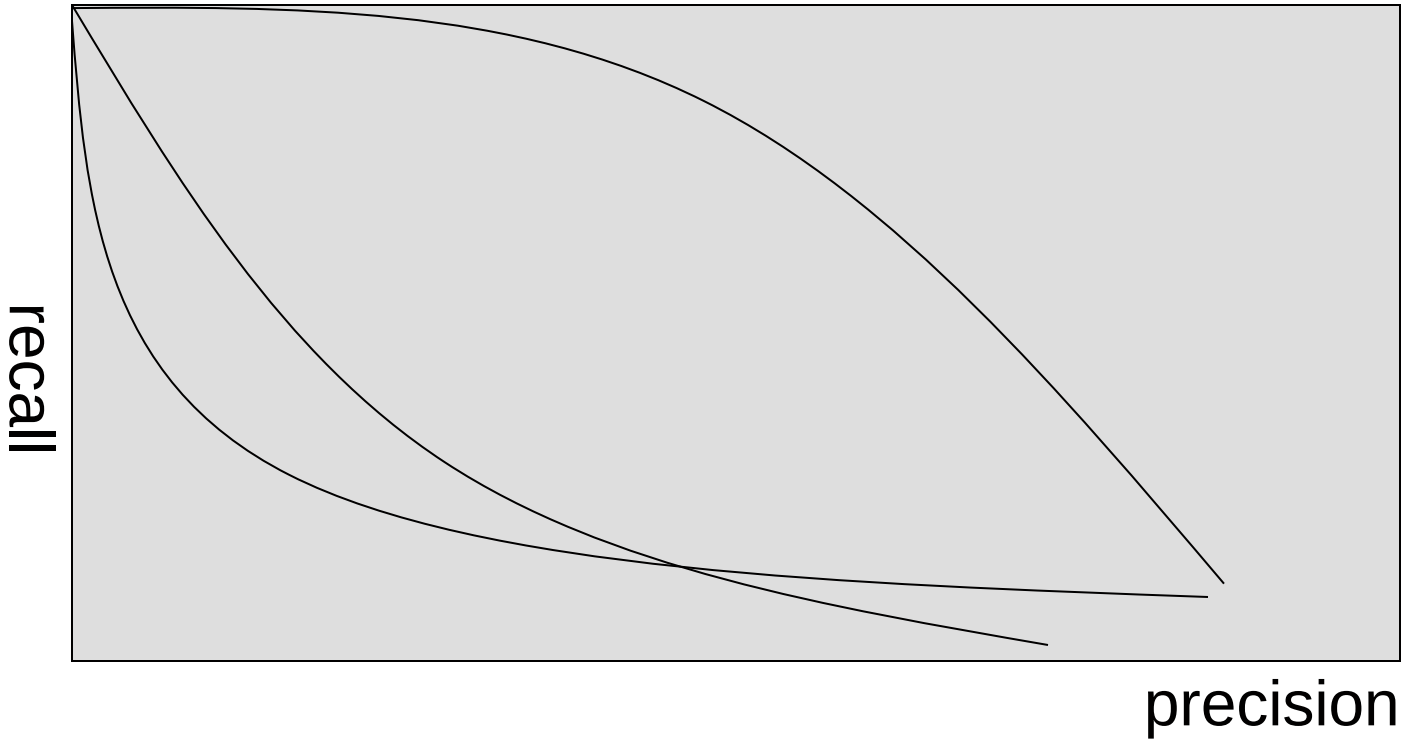
- Alternative zur ROC-Analyse.
- Stammt aus dem Information Retrieval.

- $\text{Precision} = \frac{TP}{TP + FP}$ ← Alle Instanzen mit Vorhersage „+“

- $\text{Recall} = \frac{TP}{TP + FN}$ ← Alle Instanzen mit echtem Label „+“

- Precision: P(positiv | positiv vorhergesagt)
- Recall: P(positiv vorhergesagt | ist positiv)

Precision / Recall Trade-Off



- Precision-/Recall-Kurven
- Welcher Klassifikator ist der Beste / Schlechteste

F-Measure, Breakeven Point

- Zusammenfassungen der Kurve in einer Zahl:
  - F-Measure: Harmonisches Mittel über Precision und Recall, maximiert über Schwellwert θ

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

  - Precision-Recall-Breakeven-Point: Es gibt einen Punkt θ auf der Kurve für den gilt Precision(θ) = Recall(θ) =: PRBEP