Seminar Machine Learning

Supervised Learning und Model Assessment

Alexander Bast
Dozent: Dr. Zoran Nikolić

03.05.2019

Universität zu Köln

Motivation und Einführung

Herzlich Willkommen zum Seminar Machine Learning.

Einführung

Was ist Supervised Learning?

- Die Maschine lernt durch einen "Lehrer"
- Analyse der Problemstellung
- Selektive Bereitstellung von Lerninhalten

Einführung

- Aufbereitung von vorhandenen Daten
- Bestimmung von relevanten Merkmalen
- Zusammenstellung von Datenpaaren (X,Y)

Einführung

- Eingabedaten X werden Inputs genannt
- Ausgabedaten Y werden Outputs genannt
- Outputs haben zwei Ausprägungen
- Quantitativ und Qualitativ

Motivation und Kategorisierung

Kategorisierung Überwachtes Lernen

- Binäre-Klassifikation
- Output hat 2 Ausprägungen
- Ja oder Nein, Wahr oder Falsch
- Beispiel: Spamklassifizierung

Motivation und Kategorisierung

Multiklassen-Klassifikation

- Output hat mehr als 2 Ausprägungen
- Menge an möglichen Outputs
- Beispiel: Dokumentenerkennung

Motivation und Kategorisierung

Regression

- Vorhersage eines "Trends"
- Gesucht: gute Schätzung einer zukünftigen Entwicklung
- Beispiel: Währungshandel

Motivation

Welche Zielgruppen hat das Überwachten Lernen?

- Viele Anwendungsfälle denkbar
- Unternehmen mit hohem Aufkommen an Routinearbeiten
- Unternehmen mit IT-Infrastruktur

Ausgangslage

Ausgangslage

- Kein deterministischer Zusammenhang der Daten
- Es werden nicht alle relevanten Faktoren betrachtet
- Messfehler nicht auszuschließen

Ein Lösungsansatz

- $Y=f(X)+\epsilon$ mit $E(\epsilon)=0$ unabhängig von X
- kritische Faktoren werden durch ϵ erfasst
- ermöglicht leichtere Interpretation

Bayes' Theorem

- Sei A_i, i=1,...,N eine Zerlegung der Ergebnismenge in disjunkte Ereignisse
- $P(A), P(B)$ a-priori-Wahrscheinlichkeiten
- $P(A \mid B) = \frac{P(B|A)\,P(A)}{P(B)}$ bedingte Wahrscheinlichkeit
- $P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{N} P(B|A_k)P(A_k)}$ a-posteriori-Wahrscheinlichkeit

e1071

	Class	Sex	Age	Survived	Freq
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0
7	3rd	Female	Child	No	17
8	Crew	Female	Child	No	0
9	1st	Male	Adult	No	118

e1071

```
A-priori probabilities:
Y
      No      Yes
0.676965 0.323035

Conditional probabilities:
     Class
Y             1st         2nd         3rd       Crew
  No  0.08187919 0.11208054 0.35436242 0.45167785
  Yes 0.28551336 0.16596343 0.25035162 0.29817159

     Sex
Y            Male      Female
  No  0.91543624 0.08456376
  Yes 0.51617440 0.48382560

     Age
Y           Child       Adult
  No  0.03489933 0.96510067
  Yes 0.08016878 0.91983122
```

NB_Predictions	No	Yes
No	1364	362
Yes	126	349

Eine erste Vorstellung

- θ als Gewicht in der Regression
- $f(x) = x^{\mathsf{T}}\beta$ mit $\beta = \theta$
- θ als Gewicht in der linearen Basis Expansion
- $\mathrm{f}_\theta = \sum_{k=1}^{N} h_{\mathrm{k}}(x)\theta_{\mathrm{k}}$
- mit $\mathrm{h_k(x)}= x^2{}_1, x_1 x_1{}^2, cos(x_1)$

Optimierung des Schätzers θ

- Minimierung von
- $RSS(\theta) = \sum_{k=1}^{N}(y_k - f_\theta(x_k))^2$
- Maximierung von
- $L(\theta) = \sum_{k=1}^{N} log(Pr_\theta(y_i))$

Model Assessment

Bewertungsgrundlagen

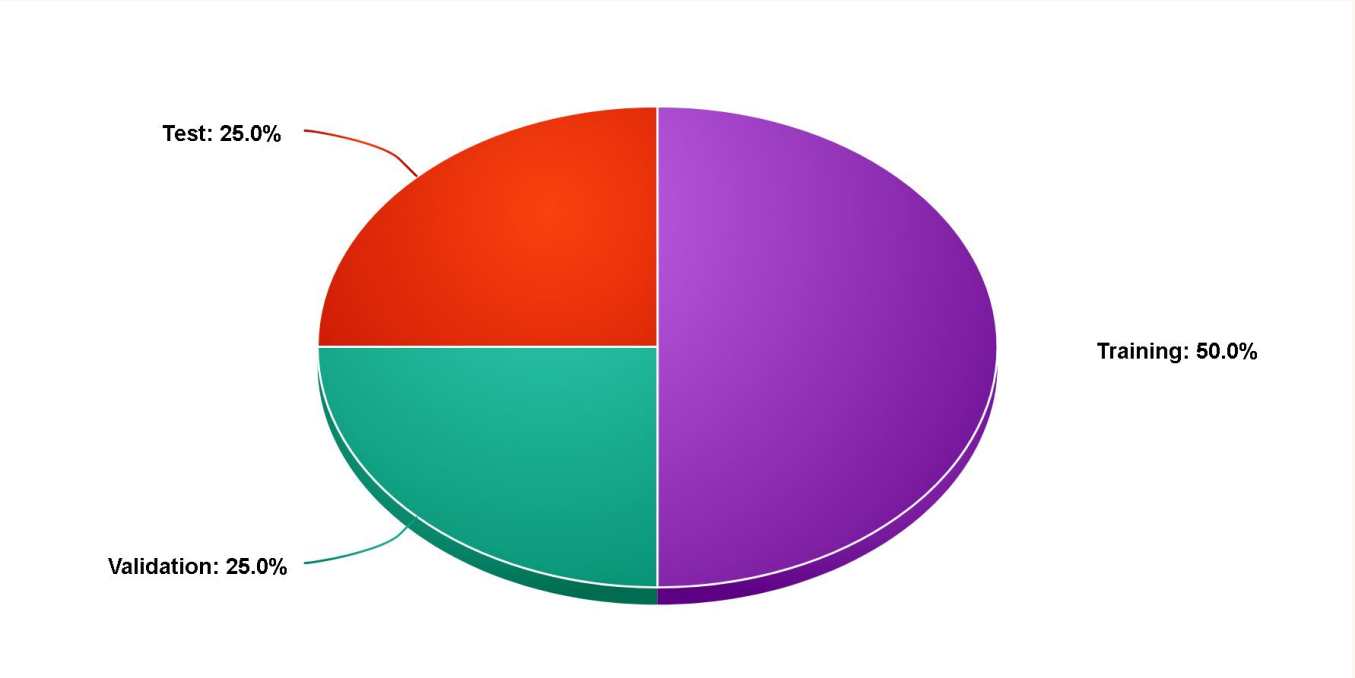
- Wodurch zeichnet sich ein gutes Modell aus?
- Antwort: Generalisierungsfähigkeit!

Bewertungsgrundlagen

- $L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{quadratischer Fehler} \\ |Y - \hat{f}(X)| & \text{absuluter Fehler} \end{cases}$
- $L(G, \hat{G}(X)) = (I(G \neq \hat{G}(X)$ (0-1 Verlust)
- $L(G, \hat{p}(X)) = -2 \sum_{k=1}^{N} I(G = k) log \hat{p}_k(X)$ (-2 x log-likelihood)

Training, Validation, Test

- Trainingsdaten
- Validationsdaten
- Testdaten



Fehlerunterscheidung

Eine Unterscheidung der Fehlerarten

- $Err_\tau = E[L(Y, \hat{f}(X))|\tau]$ Der Testfehler
- $Err = E[L(Y, \hat{f}(X))] = E[Err_\tau]$ Expected Prediction Error
- $\overline{err} = \frac{1}{N} \sum_{k=1}^{N} L(y_k, \hat{f}(x_k))$ Trainingsfehler

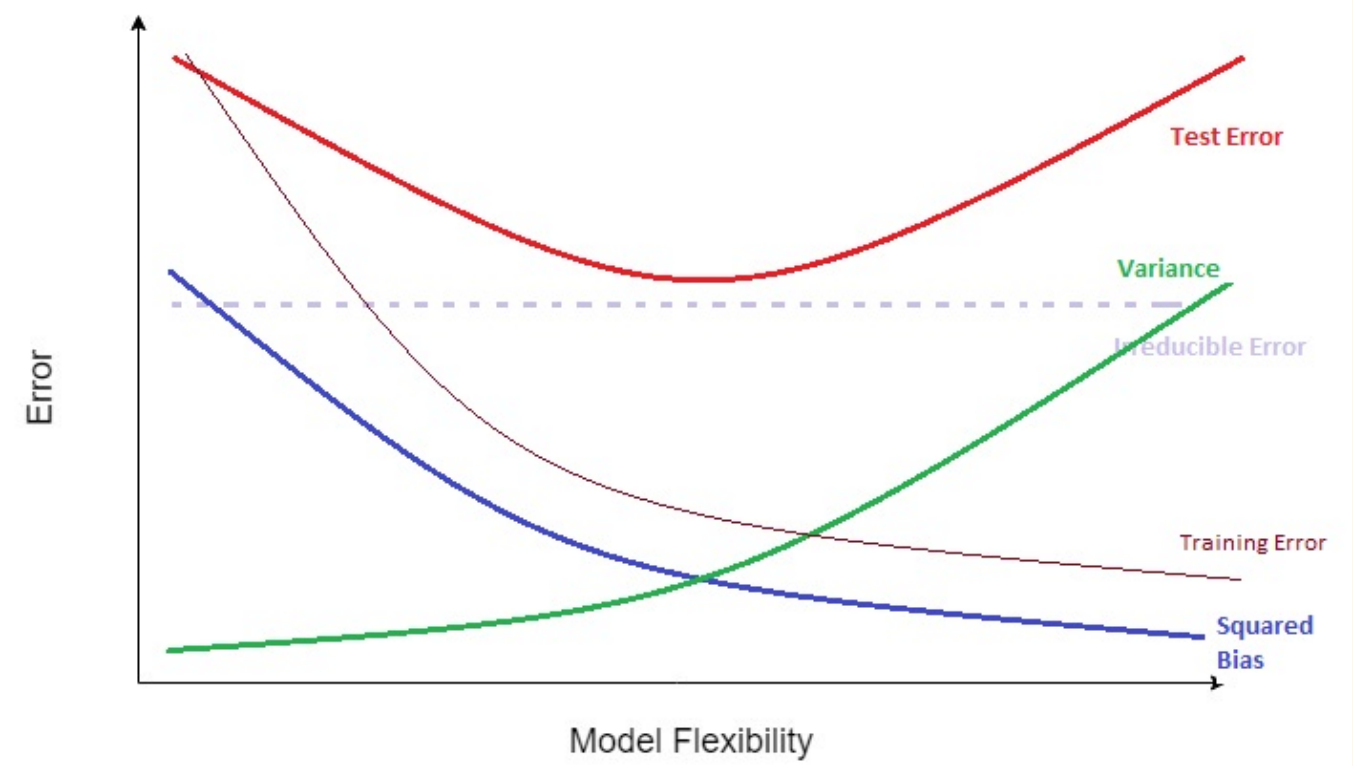
Bias

"Bias is the algorithm's tendency to consistently learn the wrong thing by not taking into account all the information in the data (underfitting)."

Varianz

"Variance is the algorithm's tendency to learn random things irrespective of the real signal by fitting highly flexible models that follow the error/noise in the data too closely (overfitting)."

Zusammenhang Bias,Varianz,Komplexität



Bias-Varianz-Dekomposition

Sei $Y = f(x) + \epsilon$, $E(\epsilon) = 0$ und σ^2. Für einen regression fit $\hat{f}(X)$ an einem Eingabepunkt x_0 kann unter Verwendung der quadratischen Verlustfunktion folgender Zusammenhang für den $Err(x_0)$ abgeleitet werden:

$$Err(x_0) = E[(Y - \hat{f}(x_0)^2 | X = x_0]$$

$$= \sigma_\epsilon^2 + [E\hat{f}(x_0 - f(x_0)]^2 + E[\hat{f}(x_0 - E\hat{f}(x_0)]^2$$

$$= \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))$$

$$= unvermeidbarerFehler + Bias^2 + Varianz$$

Bias-Varianz-Dekomposition

Für einen k-nn regression fit findet man folgende Darstellung

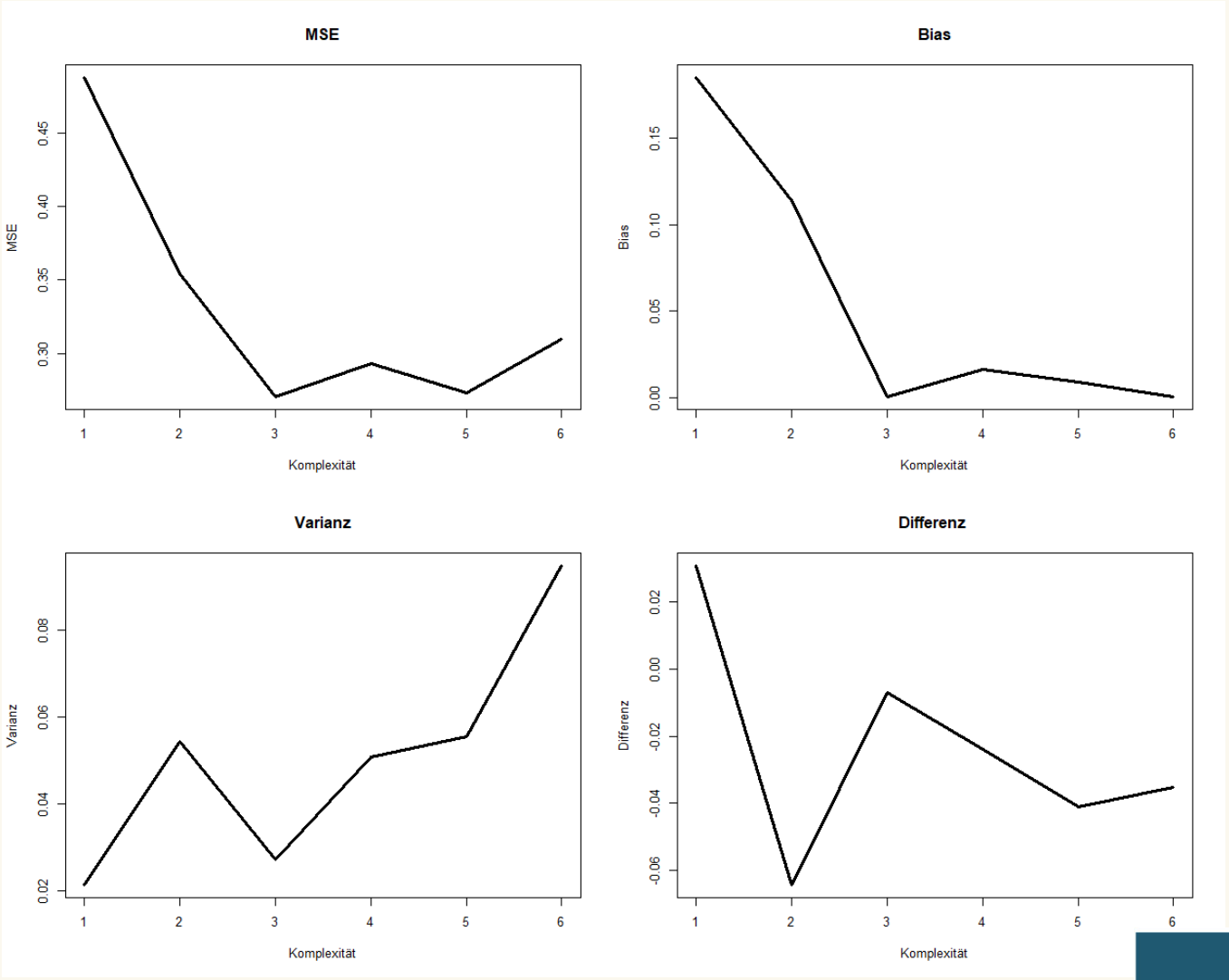
$$Err(x_0) = E[(Y - \hat{f}(x_0)^2 | X = x_0]$$

$$= \sigma_\epsilon^2 + \left[f(x_0 - \frac{1}{k} \sum_{l=1}^{k} f(x_l) \right]^2 + \frac{\sigma_\epsilon^2}{k}$$

Abschließendes Beispiel

- künstlich erzeugte Daten
- $f(x) = x + sin(2\pi x)$ an der Stelle $x_0 = 0.6$
- $\sigma = 0.5, \sigma^2 = 0.25$
- N=20, MSE als Durchschnitt aus 50 Durchläufen

Abschließendes Beispiel



Abschließendes Beispiel

	MSE	Bias	Var	Epsilon	Diff
Grad 1	0.4875873	0.1854075760	0.02136781	0.25	0.030811910
Grad 2	0.3542855	0.1142943462	0.05437759	0.25	-0.064386394
Grad 3	0.2704002	0.0002494884	0.02720678	0.25	-0.007056112
Grad 4	0.2927465	0.0159360564	0.05077899	0.25	-0.023968529
Grad 5	0.2733738	0.0088980007	0.05548879	0.25	-0.041013009
Grad 6	0.3099112	0.0003456333	0.09474335	0.25	-0.035177780

Fazit

- Supervised Learning vielseitig einsetzbar
- Erfolg abhängig von Datenaufbereitung
- Erfolg beschränkt durch Bias,Varianz und Komplexität
- Gesucht ist ein geeigneter Tradeoff zwischen drei Größen

Vielen Dank für Ihre Aufmerksamkeit!