Karlsruher Institut für Technologie

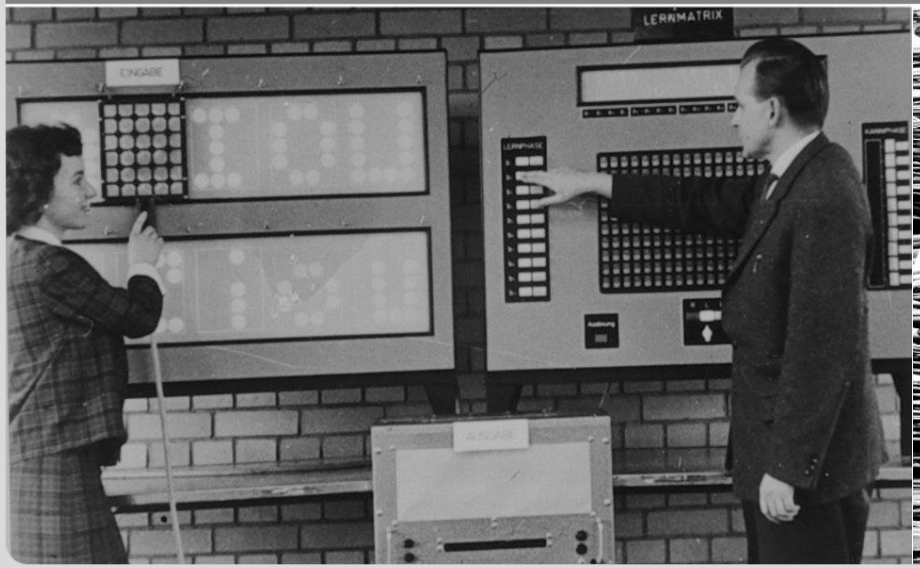
Institut für Technik der Informationsverarbeitung

Institutsleitung
Prof. Dr.-Ing. J. Becker
Prof. Dr.-Ing. E. Sax
Prof. Dr. rer. nat. W. Stork

Übung 6

Übung zu Informationstechnik II und Automatisierungstechnik – Nathalie Brenner

Prof. Dr.-Ing. Eric Sax



WIEDERHOLUNG ÜBUNG 5

Wiederholung Übung 5

Datenaufbereitung

Datenbereinigung

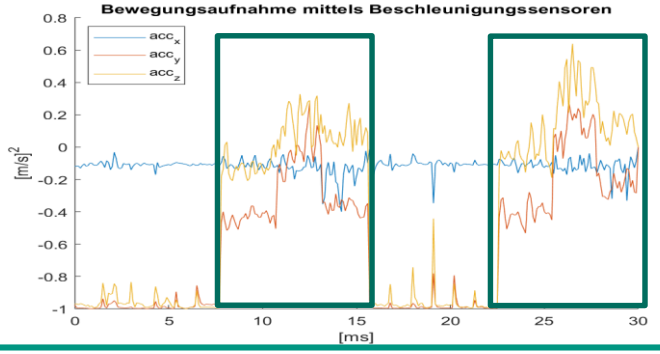
- „defekte“ Daten erkennen und bereinigen
  1. Fehlende Werte löschen/ersetzen
  2. Fensterauswahl
  3. Anomaliedetektion

Lineare Interpolation:

$$y_i = y_1 + (y_2 - y_1) * \frac{(x_i - x_1)}{(x_2 - x_1)}$$

Zeitfensterauswahl

- Tag/Nacht?
- Sommer/Winter
- Event getriggert

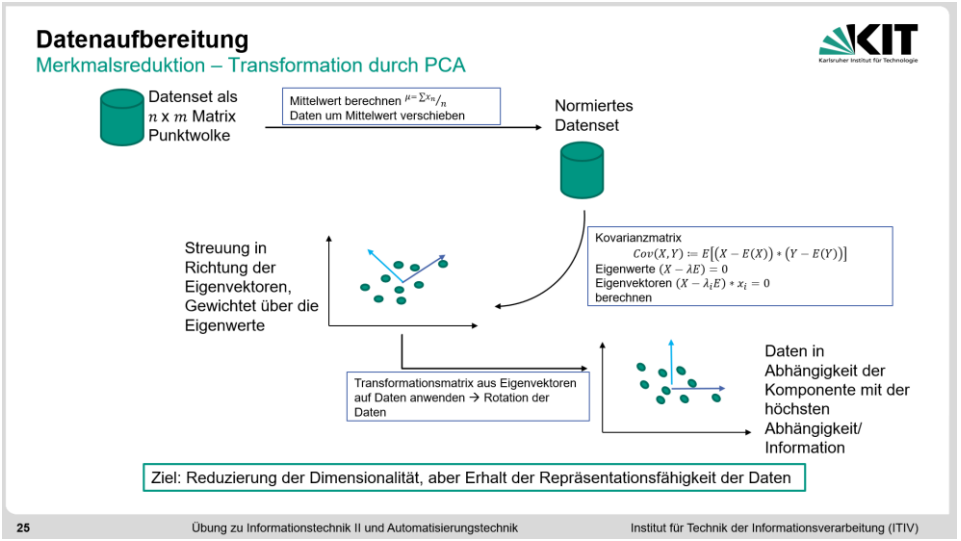


Fehlende/Fehlerhafte Werte

- Löschen
- Interpolieren
- None-Wert

```
def try_or_none(f):
    def f_or_none(x):
        try: return f(x)
        except: return None
    return f_or_none
```

Merkmalsreduktion durch PCA



Wiederholung Übung 5

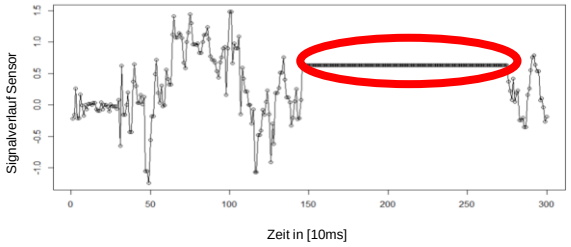
Datenaufbereitung

Datenmanipulation

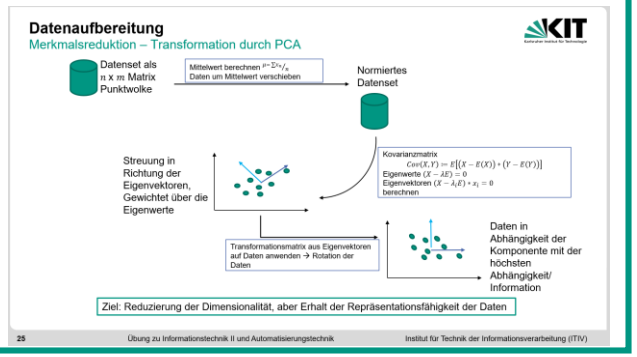
1. Umgang mit Ausreißern
2. Konvertierung der Daten nach Use Case
3. Umgang mit fehlerhaften Werten
4. Qualitätsverbesserung
5. Merkmalsreduktion

Qualitätsverbesserung

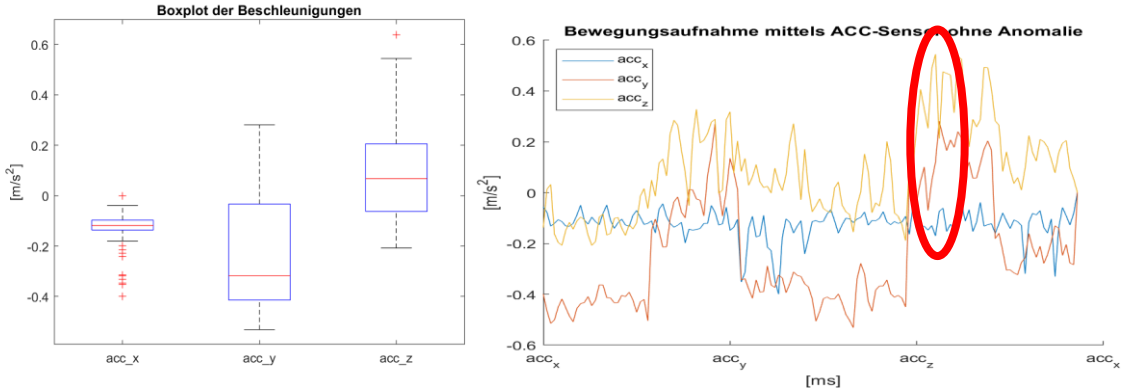
- Entfernung fehlerhafter Messungen
- Glättung von Messwerten (durch Interploation)



Merkmalsreduktion durch PCA

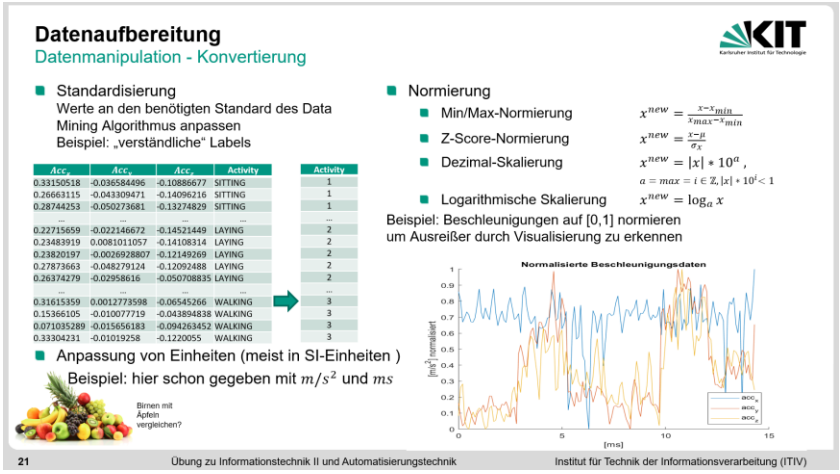


Umgang mit Ausreißern



Konvertierung

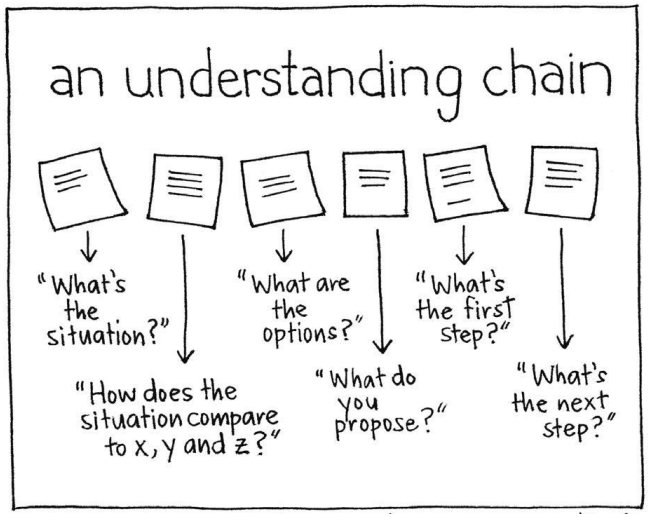
- Standardisierung
- Normierung
- Anpassung von Einheiten



INHALT ÜBUNG 6

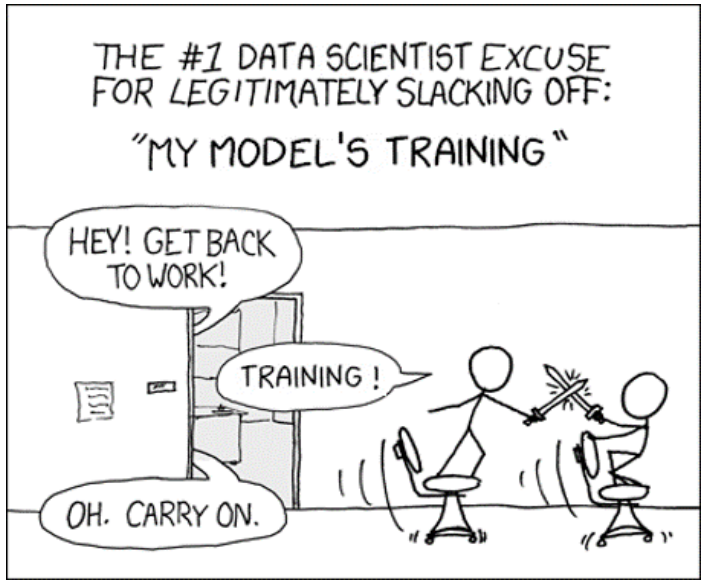
Modeling

Data Understanding ✓



Data Preparation ✓

Modeling !



Ziele der heutigen Übung

- Nach der heutigen Übung können Sie….

- …Ansätze zur Verwaltung und Analyse großer Datenbestände hinsichtlich ihrer Anwendbarkeit und Wirksamkeit einschätzen

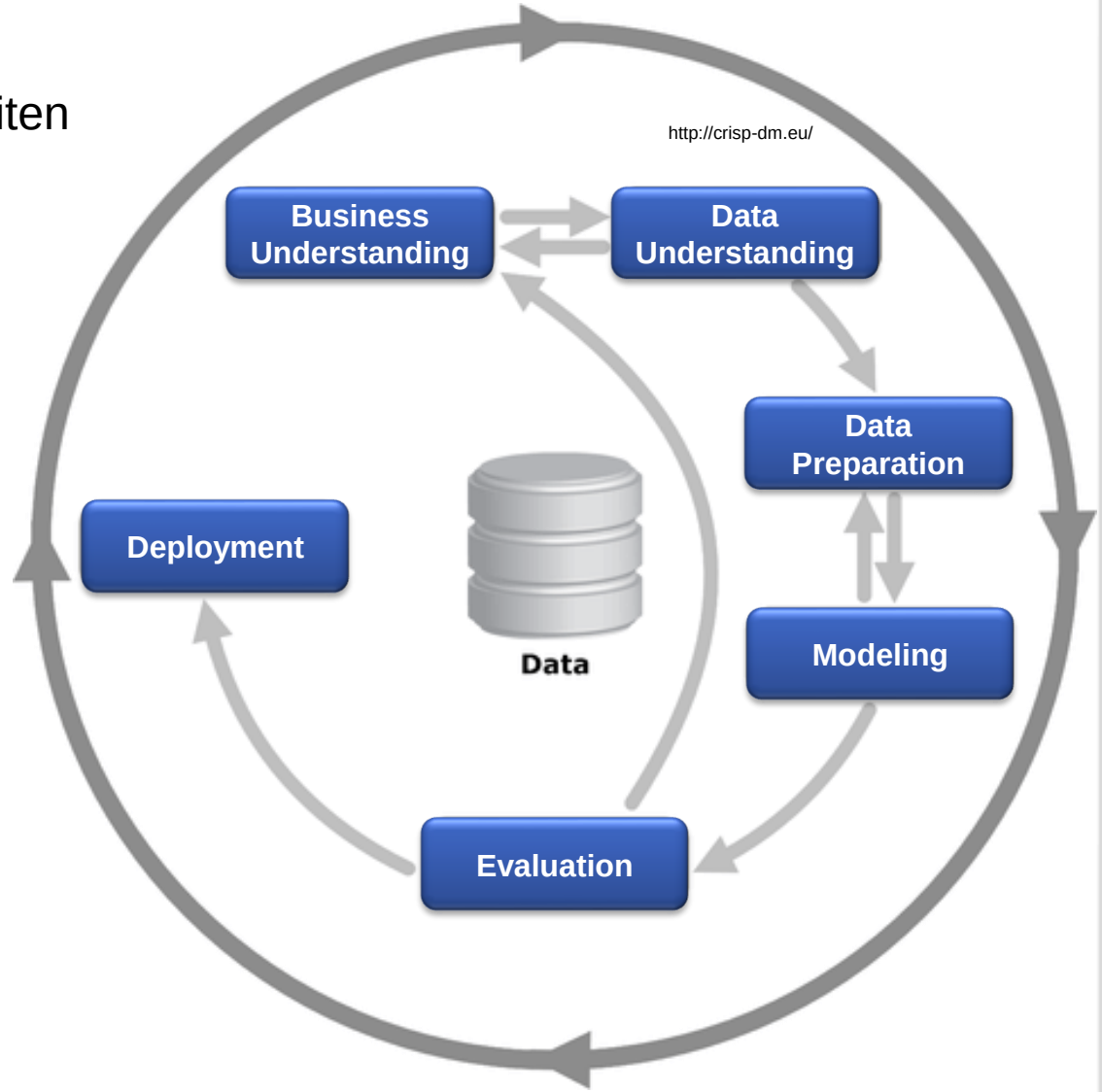
1. … Merkmale und Eigenschaften von selbstlernenden Algorithmen benennen und abgrenzen
2. … Methoden des maschinellen Lernens einordnen, beschreiben und bewerten
3. … Modelle berechnen unter Anwendung überwachter maschineller Lernmethoden
4. … Verfahren zur Auswahl einer geeigneter Methode beschreiben und anwenden

MODELING
EINFÜHRUNG IN MASCHINELLES LERNEN

Cross Industry Standard Process for Data Mining (CRISP-DM)

Modeling

- Modellbildung
  - Suche nach Modellen, Mustern oder Gesetzmäßigkeiten in den vorliegenden Daten



Algorithmus Selektion

Modeling Technique

- Group by Learning Style
  - Un- vs Supervised
  - Reinforcement
- Group by Similarity
  - Regression, Neighbor, Clustering, Decision Tree, NN, Bayes

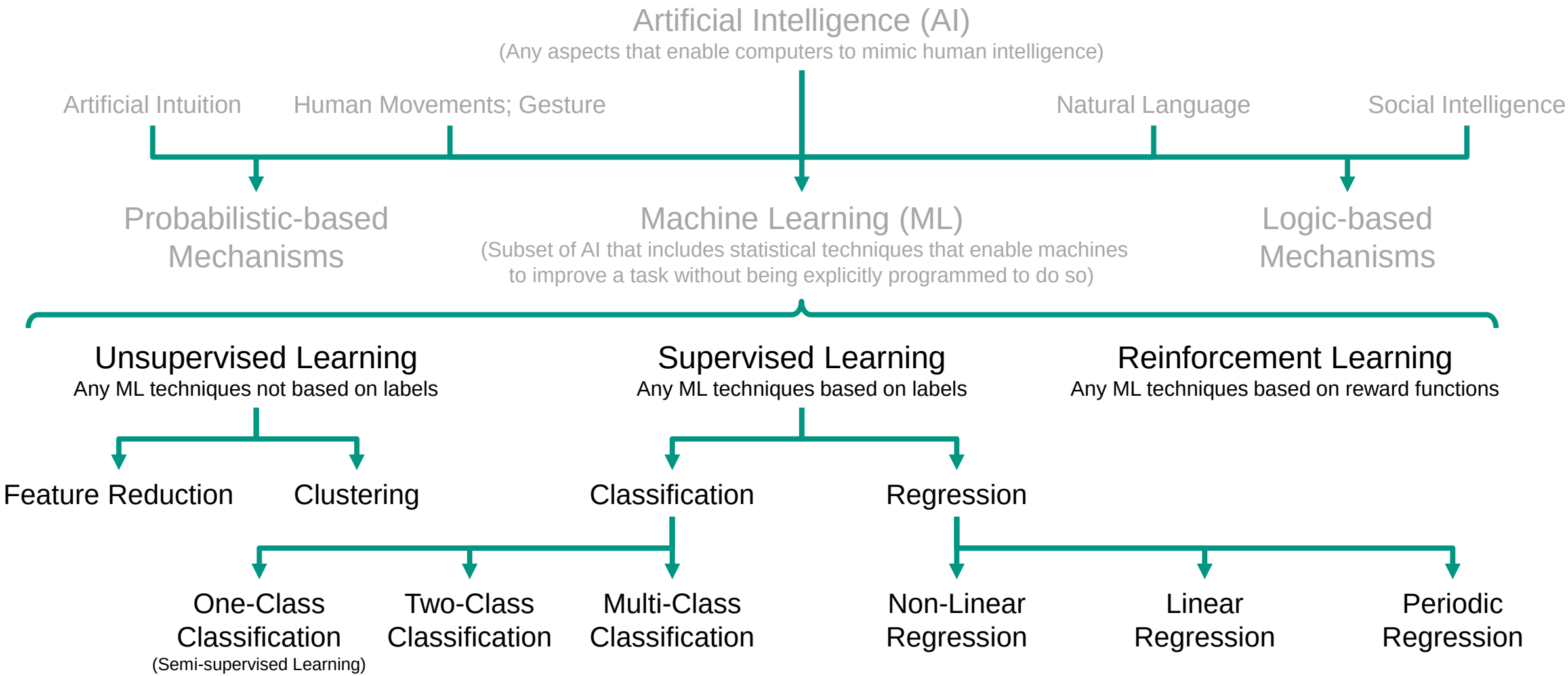
Testdesign

- Splitting
  - Test, Train, Validate

Build & Assess Model

- ROC
- Confusion Matrix
- Cross Validation
- Precision/Recall
- AUC

Modelbildung – Übersicht über Verfahren

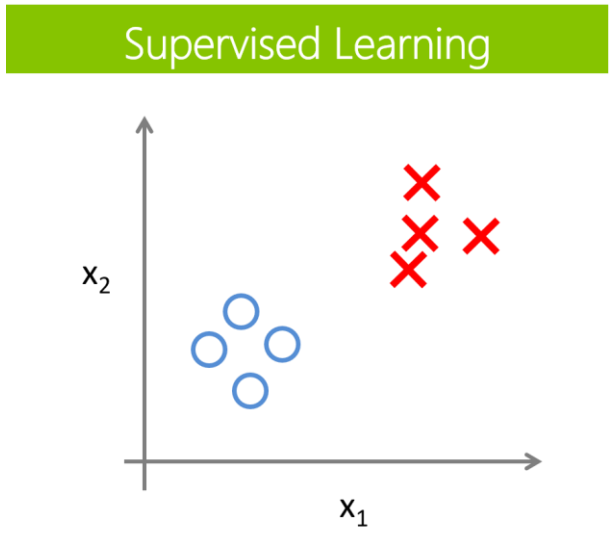


Modeling – Einteilung der Algorithmen

Group by Learning Style

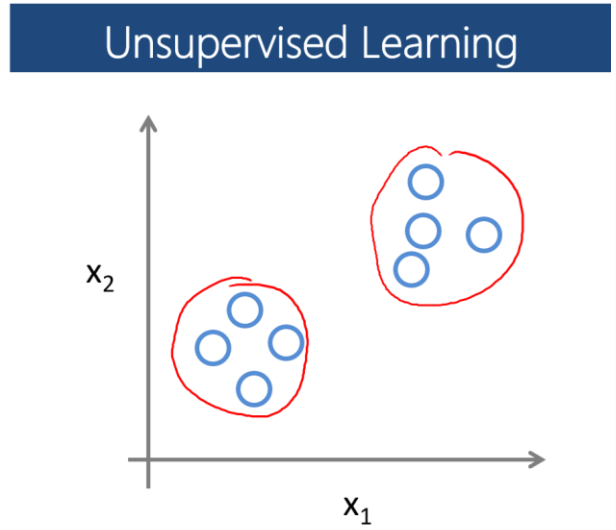
Supervised Learning

- gelabelte Daten
- Lernen/Vorhersagen von Output aus Input-Daten
- Herausforderung:
  - extrapolieren
  - generalisieren
- Beispiele
  - Klassifizierung
  - Regression



Unsupervised Learning

- ungelabelte Daten
- Auffinden von versteckten Strukturen in Daten
- Herausforderung:
  - Subjektiver als SL
  - Validierung
- Beispiele
  - Clustering
  - Dimensionsreduktion



SUPERVISED LEARNING
REGRESSION

Regression und Klassifikation

Gemeinsamkeiten und Unterschiede

Überwachtes Lernen

- Bei Trainingsdaten ist das Vorhersageattribut bekannt
- Zielgröße neuer Datensätze werden auf Basis des gelernten Modells vorhergesagt

Regressionsprobleme

- Idee
  - Bestimmung eines unbekannten **numerischen** Attributwertes (ordinal oder kategorisch durch Schwellwertsetzung)
  - Unter Benutzung beliebiger Attributwerte
- Beispiele:
  - Vorhersage von Kosten, Aufwand, etc.
  - Vorhersage von Kundenverhalten (Kündigungszeitpunkt)
  - Vorhersage zu Verkaufszahlen
  - uvm

Klassifikationsprobleme

- Idee
  - Bestimmung eines unbekannten **kategorischen** Attributwertes (ordinal mit Einschränkungen)
  - Unter Benutzung beliebiger Attributwerte
- Beispiele:
  - Klassifikation von Spam
  - Vorhersage von Kundenverhalten (Kündigung)
  - Vorhersage von Kreditwürdigkeit
  - uvm

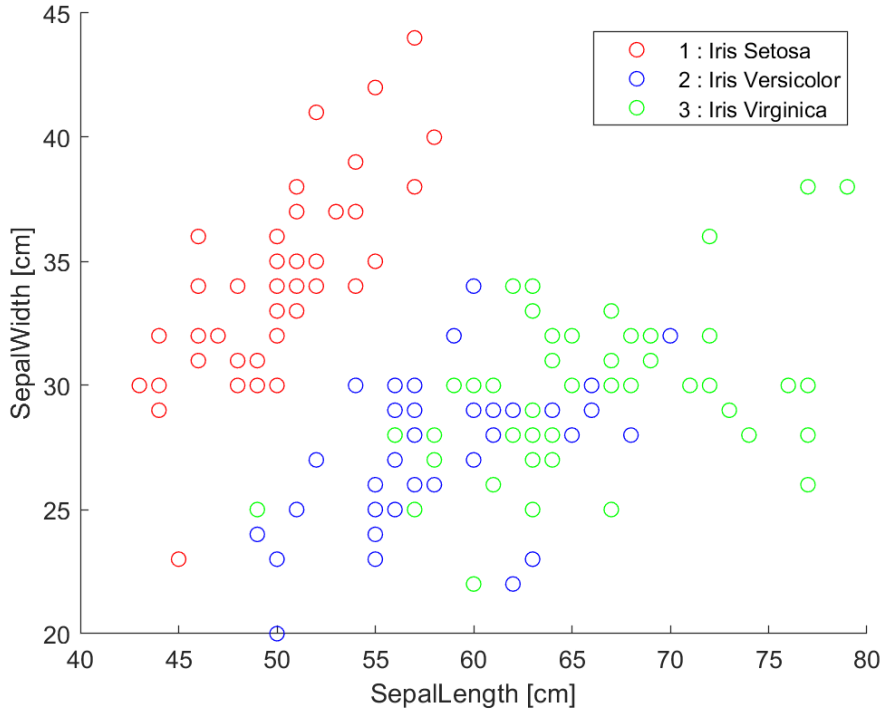
Regression

Berechnung der Linearen Regression

Regressionsgerade $y = \frac{S_{xy}}{S_{xx}} x + \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$

$$S_{xy} = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$$

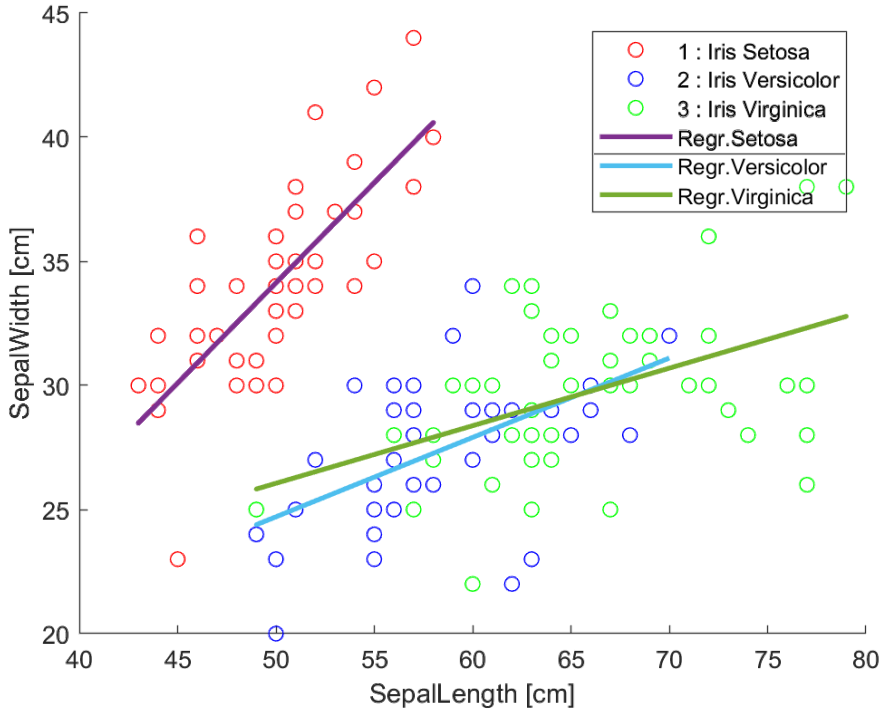
$$S_{xx} = \big((x_1 - \bar{x}) * (y_1 - \bar{y})\big) + \big((x_2 - \bar{x}) * (y_2 - \bar{y})\big) + \big((x_n - \bar{x}) * (y_n - \bar{y})\big)$$



Regressionsgerade:
$y = 0.8072x - 6.2301$

Regressionsgerade:
$y = 0.3197x + 8.7215$

Regressionsgerade:
$y = 0.2319x + 14.4631$



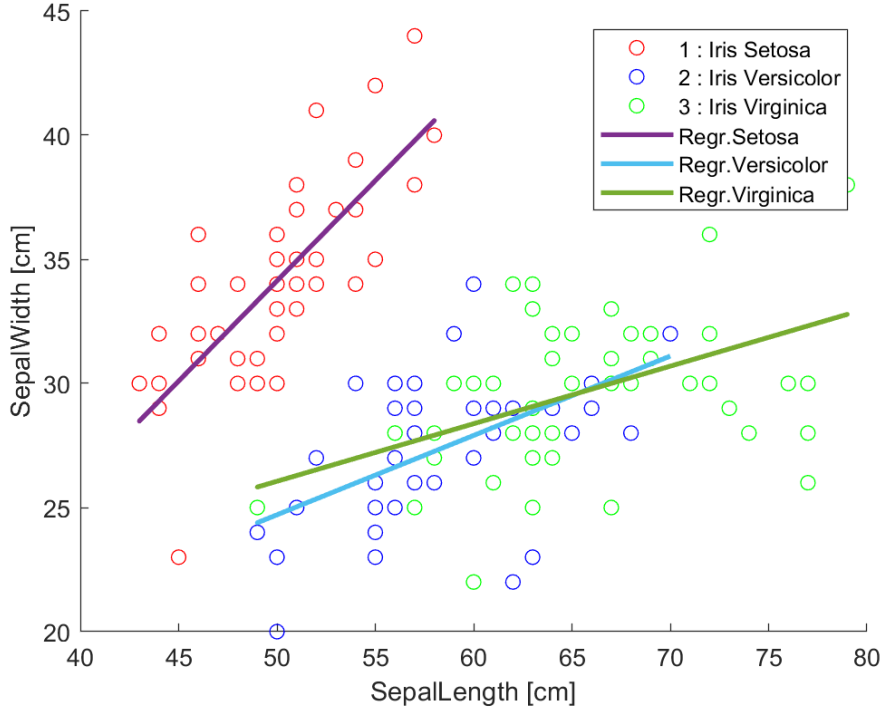
Regression
Quadratische Regression

Regressionsgerade $y = ax^2 + bx + c$

$$a = \frac{\left(\overline{y_i x_i^2} - \overline{y_i} \cdot \overline{x_i^2}\right) \cdot \left(\overline{x_i^2} - \left(\overline{x_i}\right)^2\right) - \left(\overline{y_i x_i} - \overline{y_i} \cdot \overline{x_i}\right) \cdot \left(\overline{x_i^3} - \overline{x_i} \cdot \overline{x_i^2}\right)}{\left(\overline{x_i^4} - \left(\overline{x_i^2}\right)^2\right) \cdot \left(\overline{x_i^2} - \left(\overline{x_i}\right)^2\right) - \left(\overline{x_i^3} - \overline{x_i} \cdot \overline{x_i^2}\right)^2}$$

$$b = \frac{\overline{y_i x_i} - \overline{y_i} \cdot \overline{x_i} - a \cdot \left(\overline{x_i^3} - \overline{x_i} \cdot \overline{x_i^2}\right)}{\overline{x_i^2} - \left(\overline{x_i}\right)^2}$$

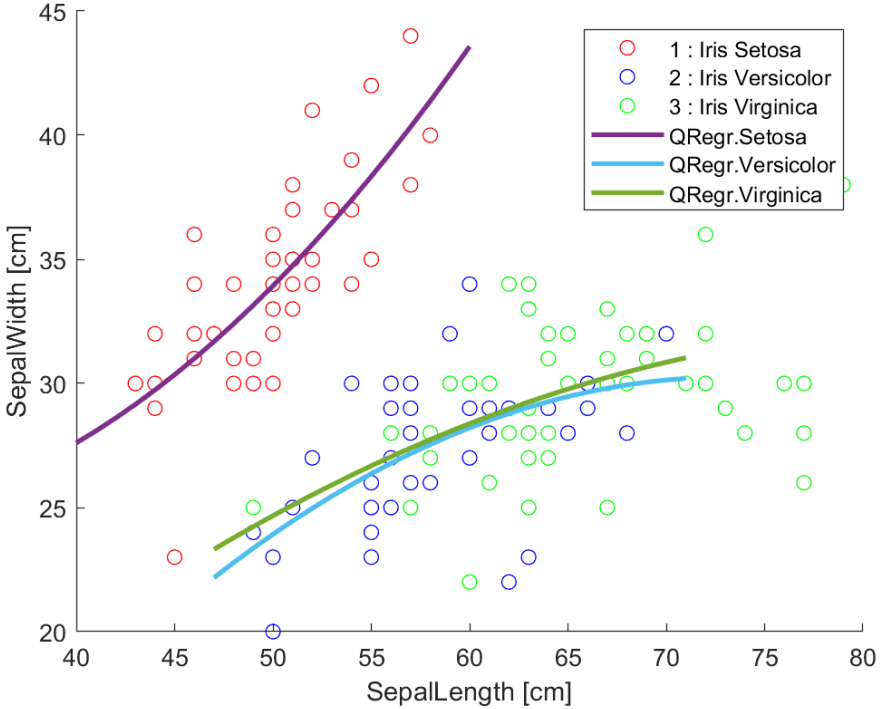
$$c = \overline{y_i} - a \cdot \overline{x_i^2} - b \cdot \overline{x_i}$$



Regressionskurve:
$y = 0.0166x^2 - 0.8598x + 35.4623$

Regressionskurve:
$y = -0.0118x^2 - 1.7305x - 33.0252$

Regressionskurve:
$y = -0.0062x^2 + 1.0516x - 12.4383$

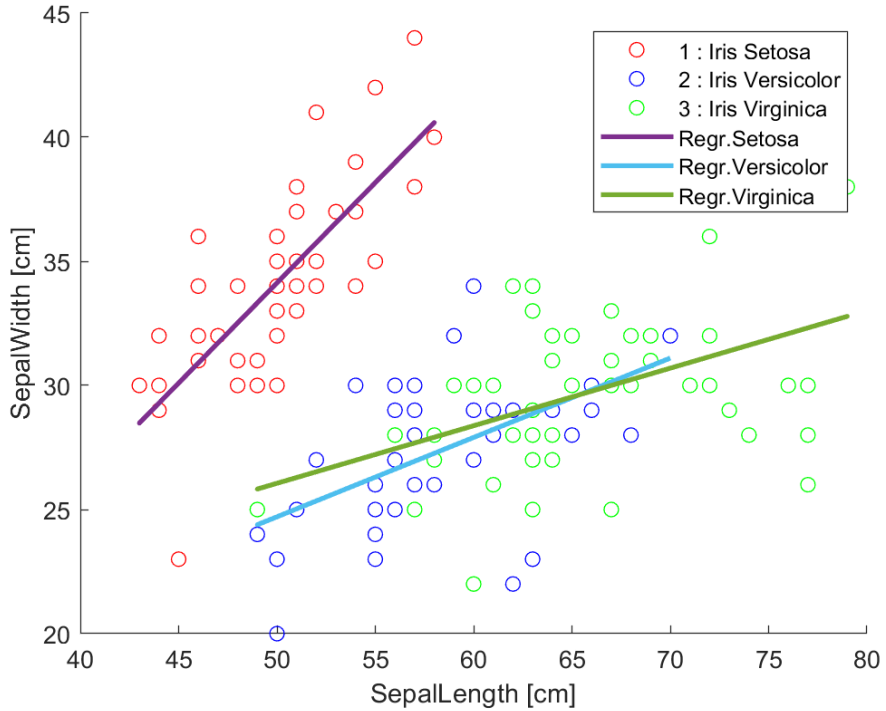


Regression
Methodenauswahl

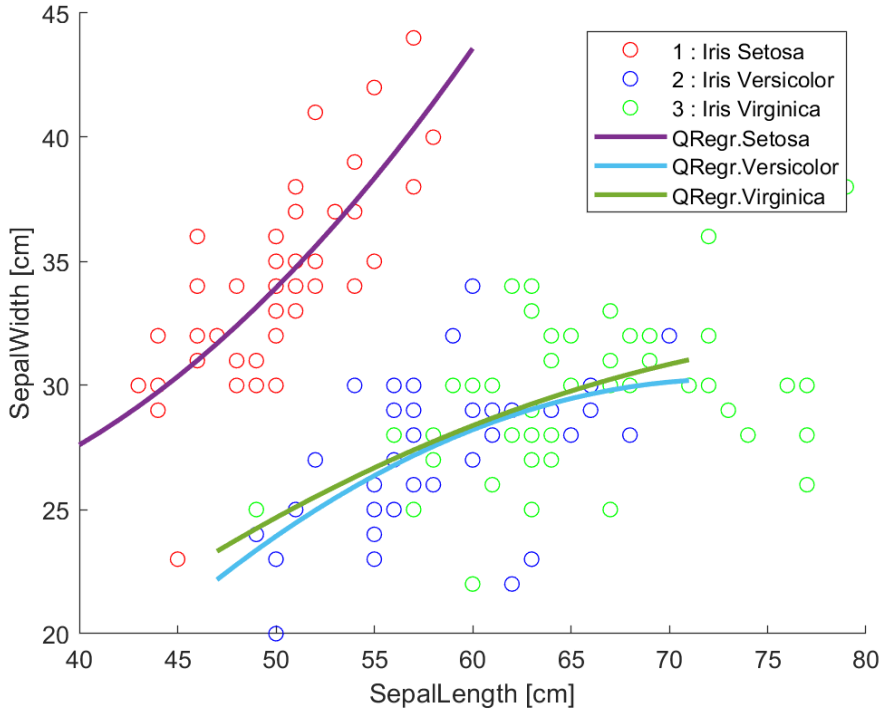
Es sei ein neuer Datenpunkt für die „Iris-Analyse“ gegeben

Wie wird nun entschieden, welche Methode sich am besten eignet um eine Vorhersage zu treffen, zu welcher Klasse der neue Punkt gehört?

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	51	35	14	2	Iris-setosa
2	49	30	14	2	Iris-setosa
3	47	32	13	2	Iris-setosa
4	46	31	15	2	Iris-setosa
5	50	36	14	2	Iris-setosa
51	70	32	47	14	Iris-versicolor
52	64	32	45	15	Iris-versicolor
53	69	31	49	15	Iris-versicolor
54	55	23	40	13	Iris-versicolor
55	65	28	46	15	Iris-versicolor
152	**52**	**50**	**30**	**6**	**?????**



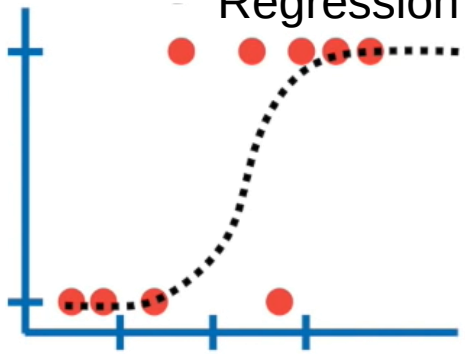
k-fold Cross Validation



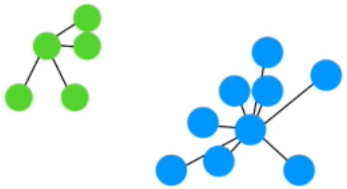
Supervised Learning

Methodenauswahl – k-fold Cross Validation

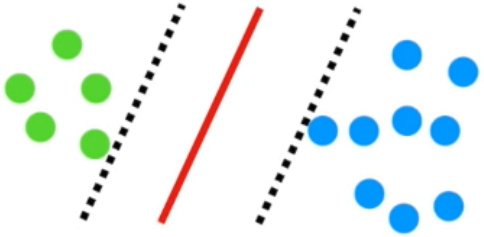
Regression



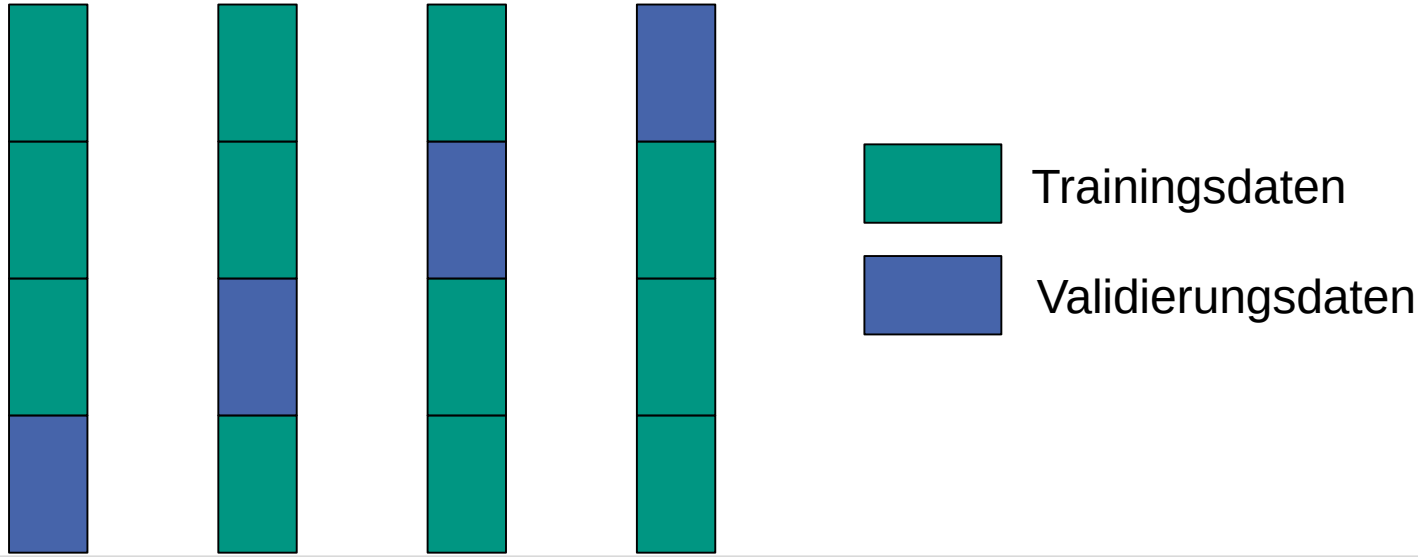
K-nearest neighbors



Support Vector machines



- Kompletten Datensatz teilen in 70% Trainings- und 30% Testdaten
  1. Die Trainingsdaten werden in k Teile unterteilt
  2. Auf Basis von k-1 Teilen wird das Modell berechnet und im Anschluss mit dem verbleibenden Teil validiert.
  3. Wiederholen bis alle Teile einmalig der Validierung dienten
  4. MSE berechnen und merken
  5. Wiederholung der Punkte 1.-4. für jede Methode
  6. Auswahl der Methode mit dem geringsten MSE
- Modellbildung auf Basis der gesamten 70% anhand der ausgewählten Methode

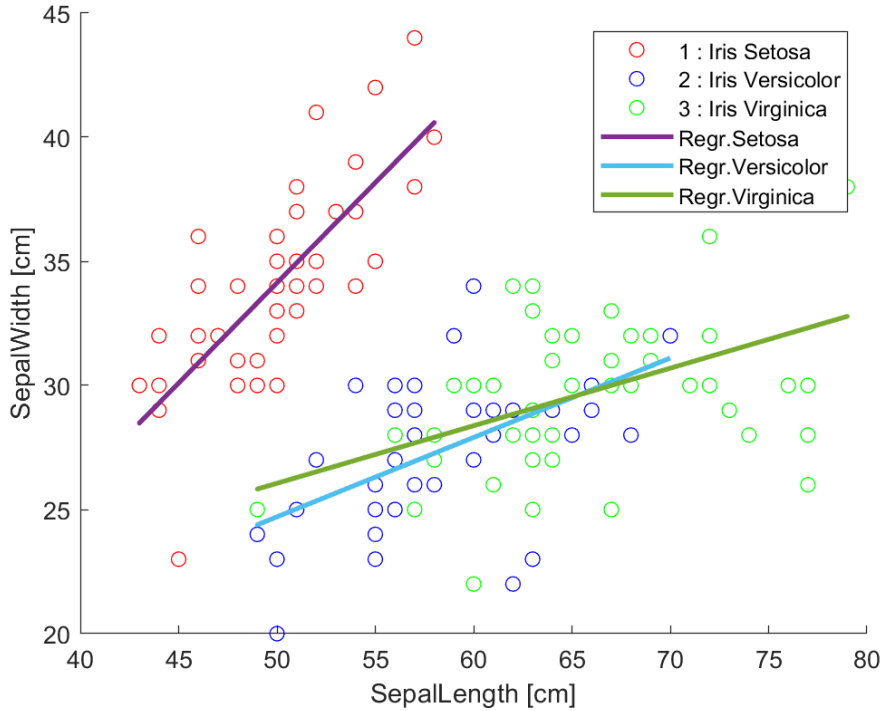


Supervised Learning

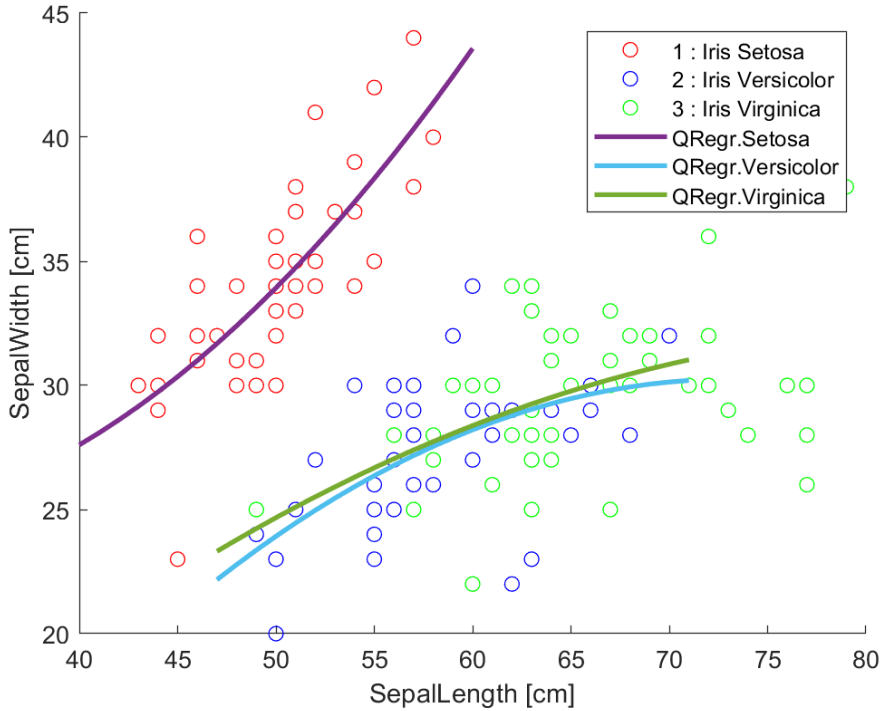
Methodenauswahl – k-fold Cross Validation

Mean Squared Error $\quad MSE = \frac{1}{n}\sum\left(y(x) - y^*(x)\right)^2$

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	51	35	14	2	Iris-setosa
2	49	30	14	2	Iris-setosa
3	47	32	13	2	Iris-setosa
4	46	31	15	2	Iris-setosa
5	50	36	14	2	Iris-setosa
51	70	32	47	14	Iris-versicolor
52	64	32	45	15	Iris-versicolor
53	69	31	49	15	Iris-versicolor
54	55	23	40	13	Iris-versicolor
55	65	28	46	15	Iris-versicolor
152	**52**	**50**	**30**	**6**	**?????**



Iris Setosa	
Regr.art	MSE
L. Regr	6.5554
Q. Regr.	6.2262

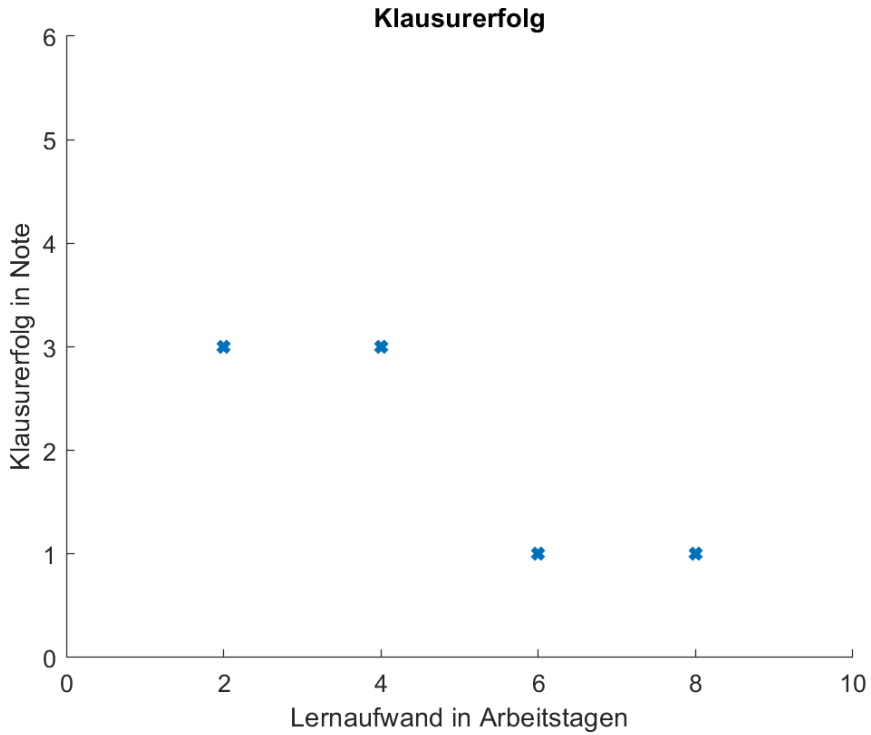


Regression

Zwischenübung

Es ist der Lernaufwand x_i von vier Personen, sowie der Klausurerfolg y_i(in Note) gegeben. Berechnen Sie die Regressionsgerade $\mathbf{y} = \frac{\boldsymbol{s_{xy}}}{\boldsymbol{s_{xx}}}\boldsymbol{x} + \overline{\boldsymbol{y}} - \frac{\boldsymbol{s_{xy}}}{\boldsymbol{s_{xx}}}\overline{\boldsymbol{x}}$

x_i	y_i
8	1
4	3
6	1
2	3
$\bar{x}$=5	$\bar{y}$=2

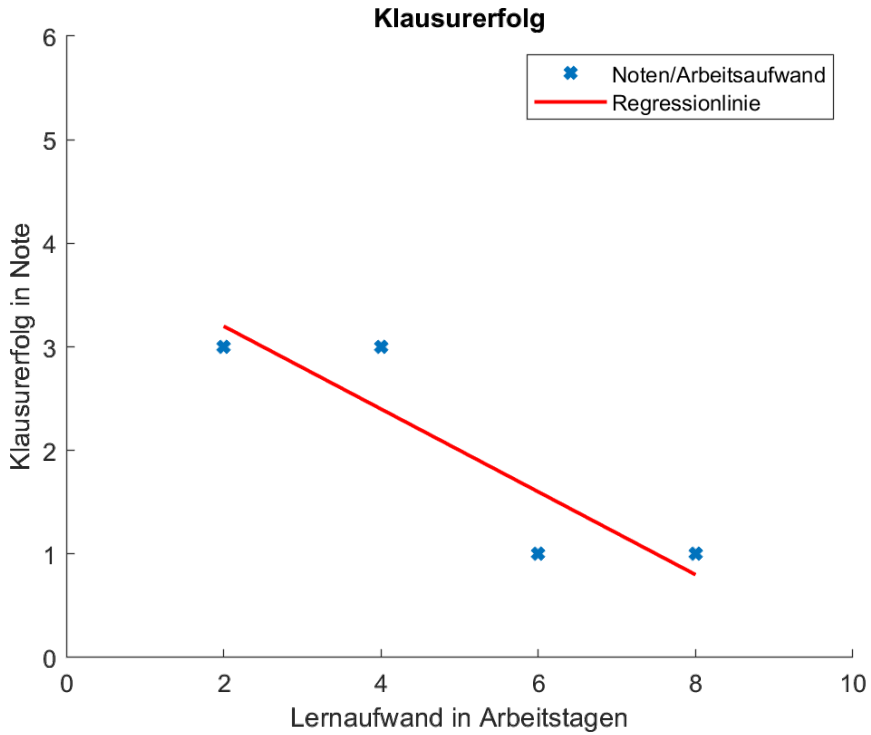


Regression

Zwischenübung - Lsg

Es ist der Lernaufwand x_i von vier Personen, sowie der Klausurerfolg y_i(in Note) gegeben. Berechnen Sie die Regressionsgerade $\mathbf{y} = \frac{S_{xy}}{S_{xx}}\boldsymbol{x} + \bar{\boldsymbol{y}} - \frac{S_{xy}}{S_{xx}}\bar{\boldsymbol{x}}$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})\,(y_i - \bar{y})$
8	1					
4	3					
6	1					
2	3					
$\bar{x}$=5	$\bar{y}$=5					



SUPERVISED LEARNING
KLASSIFIKATION

Regression und Klassifikation

Gemeinsamkeiten und Unterschiede

Überwachtes Lernen

- Bei Trainingsdaten ist das Vorhersageattribut bekannt
- Zielgröße neuer Datensätze werden auf Basis des gelernten Modells vorhergesagt

Regressionsprobleme

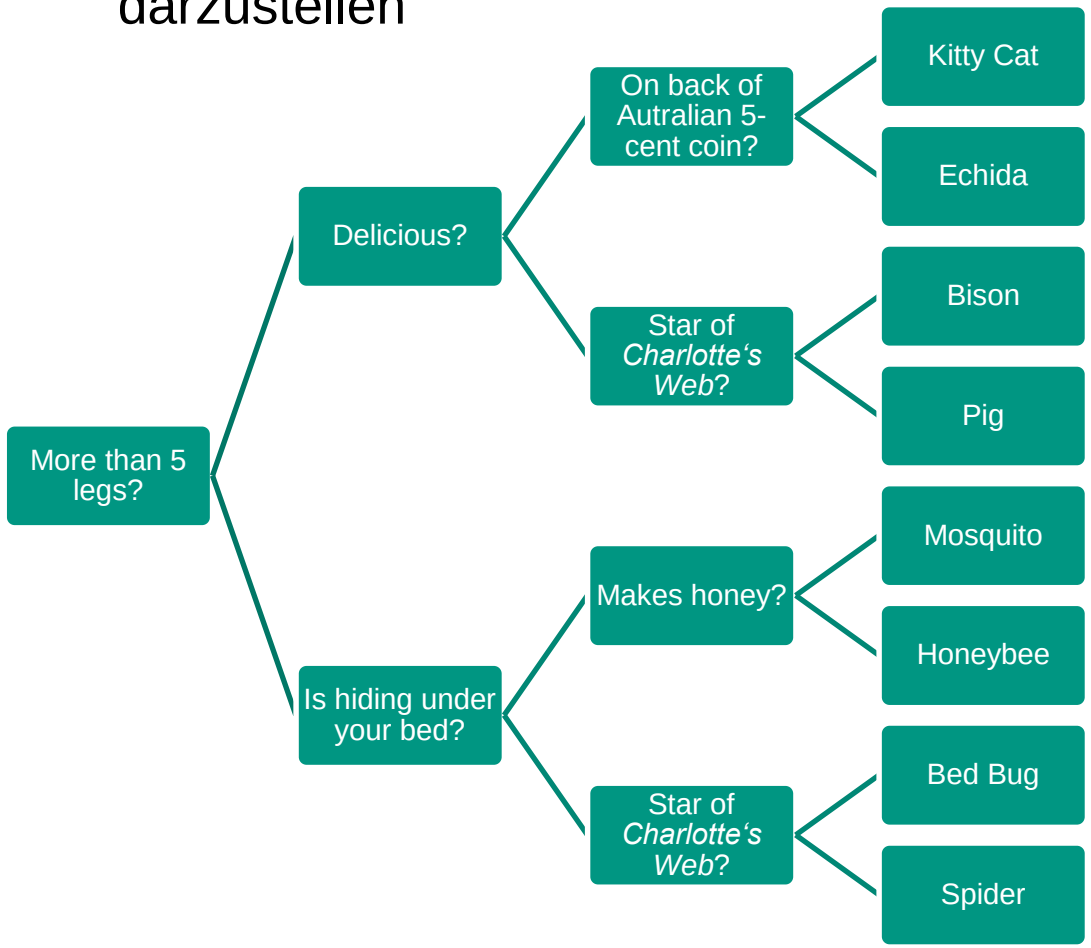
- Idee
  - Bestimmung eines unbekannten **numerischen** Attributwertes (ordinal oder kategorisch durch Schwellwertsetzung)
  - Unter Benutzung beliebiger Attributwerte
- Beispiele:
  - Vorhersage von Kosten, Aufwand, etc.
  - Vorhersage von Kundenverhalten (Kündigungszeitpunkt)
  - Vorhersage zu Verkaufszahlen
  - uvm

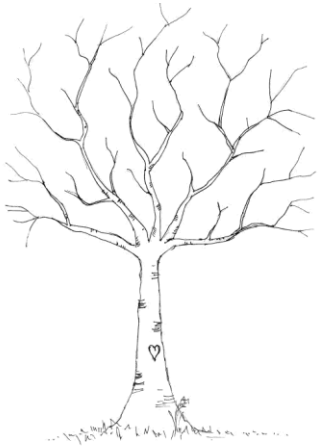
Klassifikationsprobleme

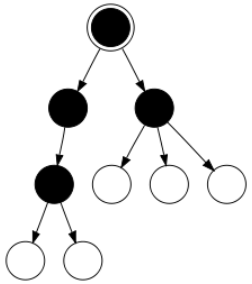
- Idee
  - Bestimmung eines unbekannten **kategorischen** Attributwertes (ordinal mit Einschränkungen)
  - Unter Benutzung beliebiger Attributwerte
- Beispiele:
  - Klassifikation von Spam
  - Vorhersage von Kundenverhalten (Kündigung)
  - Vorhersage von Kreditwürdigkeit
  - uvm

Klassifikation
Decision Tree

- Ein Decision Tree verwendet eine Baumstruktur, um eine Reihe möglicher Entscheidungspfade und das Ergebnis für jeden Pfad darzustellen





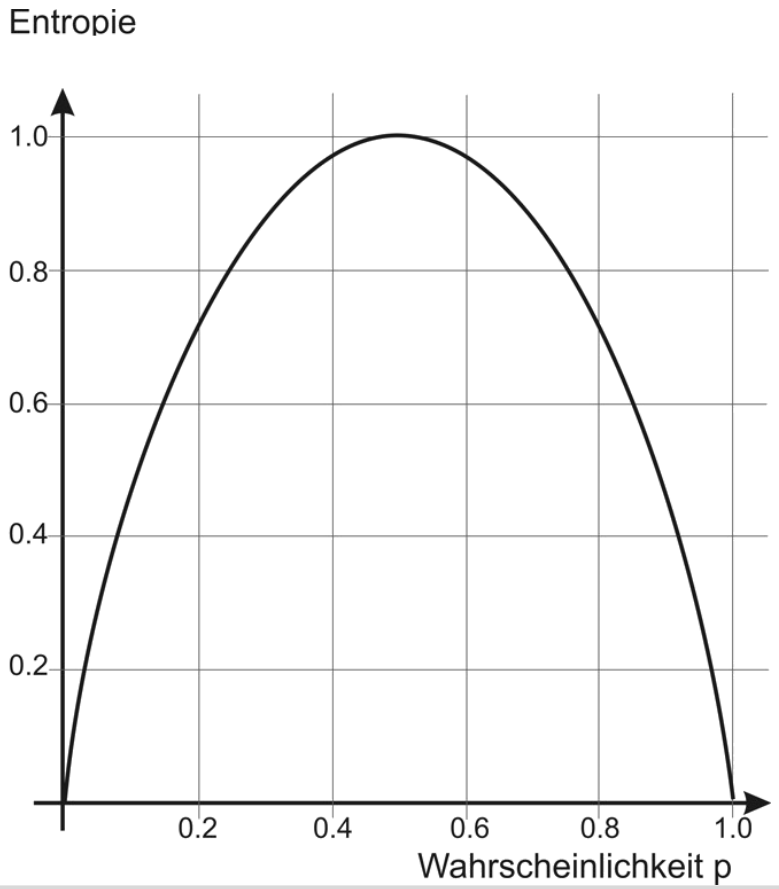


Aufbau?

Klassifikation

Decision Tree - Entwurf

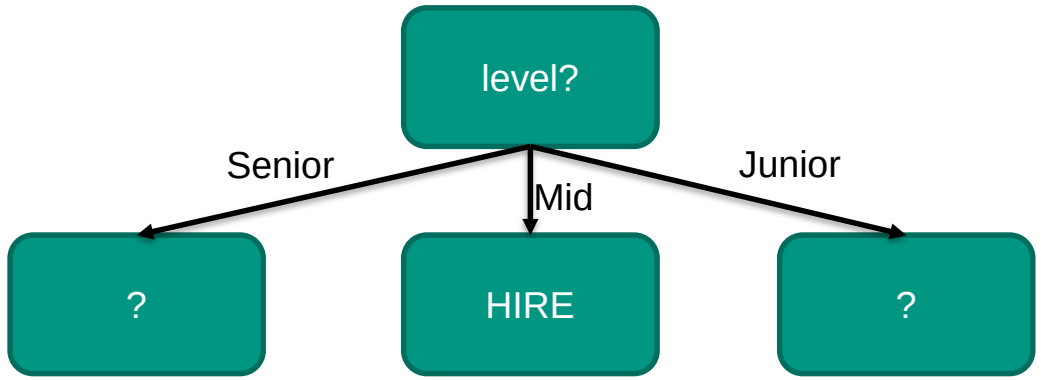
- Für den Entwurf des Decision Tree muss entschieden werden
  - Welche „Fragen“ werden an die Daten gestellt (*Auswahl der genutzten Attribute)*
  - In welcher Reihenfolge werden die „Fragen“ an die Daten gestellt *(Reihenfolge der Attributsabfrage)*
- ➢ Abhilfe schafft Entropie
  - Entropie $\mathrm{H}(\boldsymbol{S})$: Aussage über Informationsgehalt
    $\mathbf{H}(\boldsymbol{S}) = -\boldsymbol{p_1}\,\mathbf{log}\,\boldsymbol{p_1} - \boldsymbol{p_2}\,\boldsymbol{log}\,\boldsymbol{p_2} - \cdots - \boldsymbol{p_n}\,\boldsymbol{log}\,\boldsymbol{p_n}$
    mit p_i als Wahrscheinlich des Ereignis i
  - Je eher p_i gegen 0 oder 1 geht
    desto $H(S_i) \rightarrow 0$



Decision Tree Entwurf

„Should you hire the candidate?“

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
level	Senior	Senior	Mid	Junior	Junior	Junior	Mid	Senior	Senior	Junior	Senior	Mid	Mid	Junior
language	Java	Java	Python	Python	R	R	R	Python	R	Python	Python	Python	Java	Python
tweets	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
PhD	No	Yes	No	No	No	Yes	Yes	No	No	No	Yes	Yes	No	Yes
Hire?	False	False	True	True	True	False	True	False	True	True	True	True	True	False



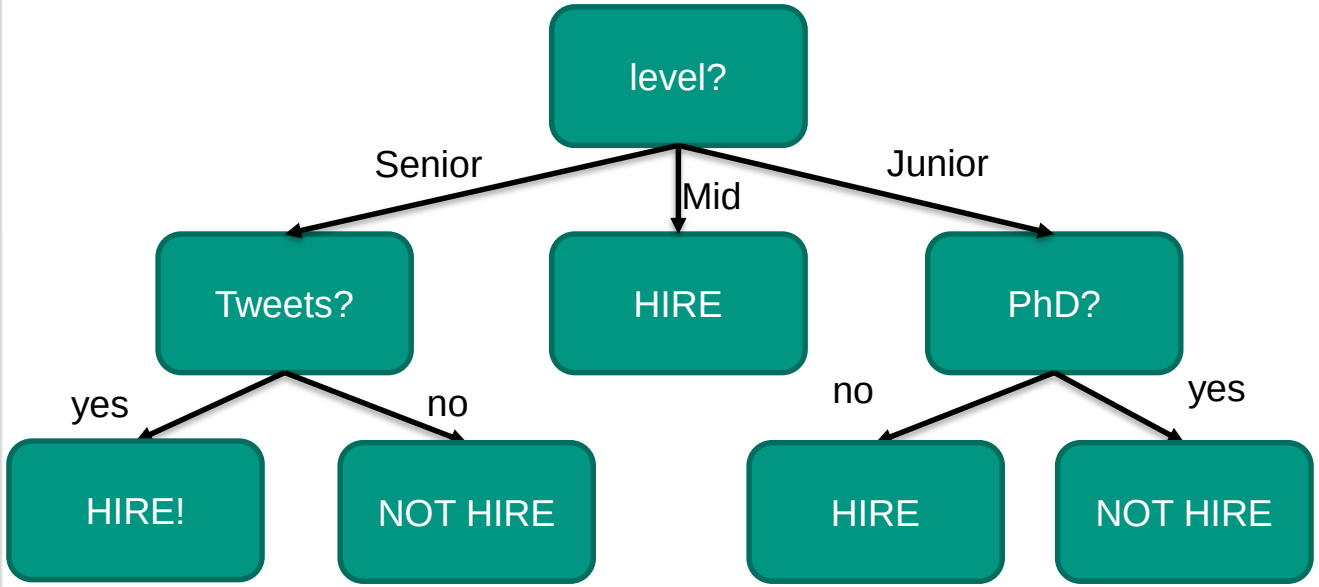
Entropie:

level:	0.693536
language:	0.860131
tweets:	0.788450
PhD:	0.892158

Decision Tree Entwurf

„Should you hire the candidate?“

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
level	Senior	Senior	Mid	Junior	Junior	Junior	Mid	Senior	Senior	Junior	Senior	Mid	Mid	Junior
language	Java	Java	Python	Python	R	R	R	Python	R	Python	Python	Python	Java	Python
tweets	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No
PhD	No	Yes	No	No	No	Yes	Yes	No	No	No	Yes	Yes	No	Yes
Hire?	False	False	True	True	True	False	True	False	True	True	True	True	True	False



Entropie:

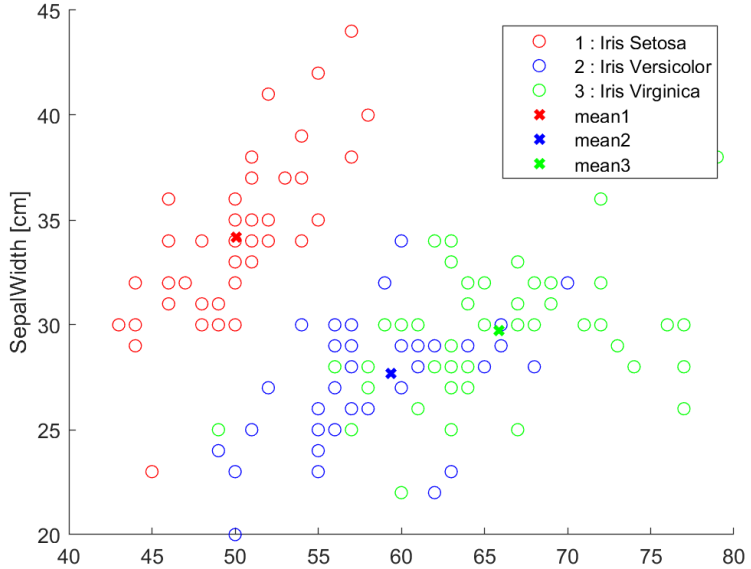
level:	0.693536
language:	0.860131
tweets:	0.788450
PhD:	0.892158

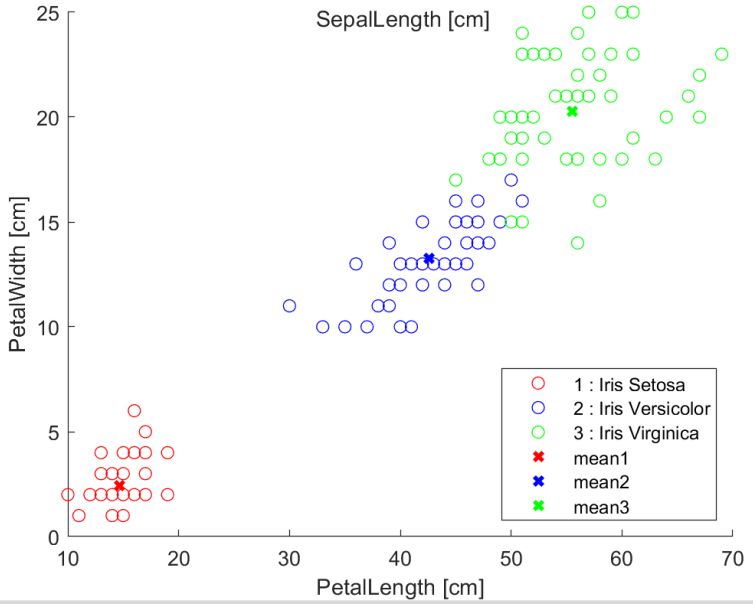
Entropie „Senior“

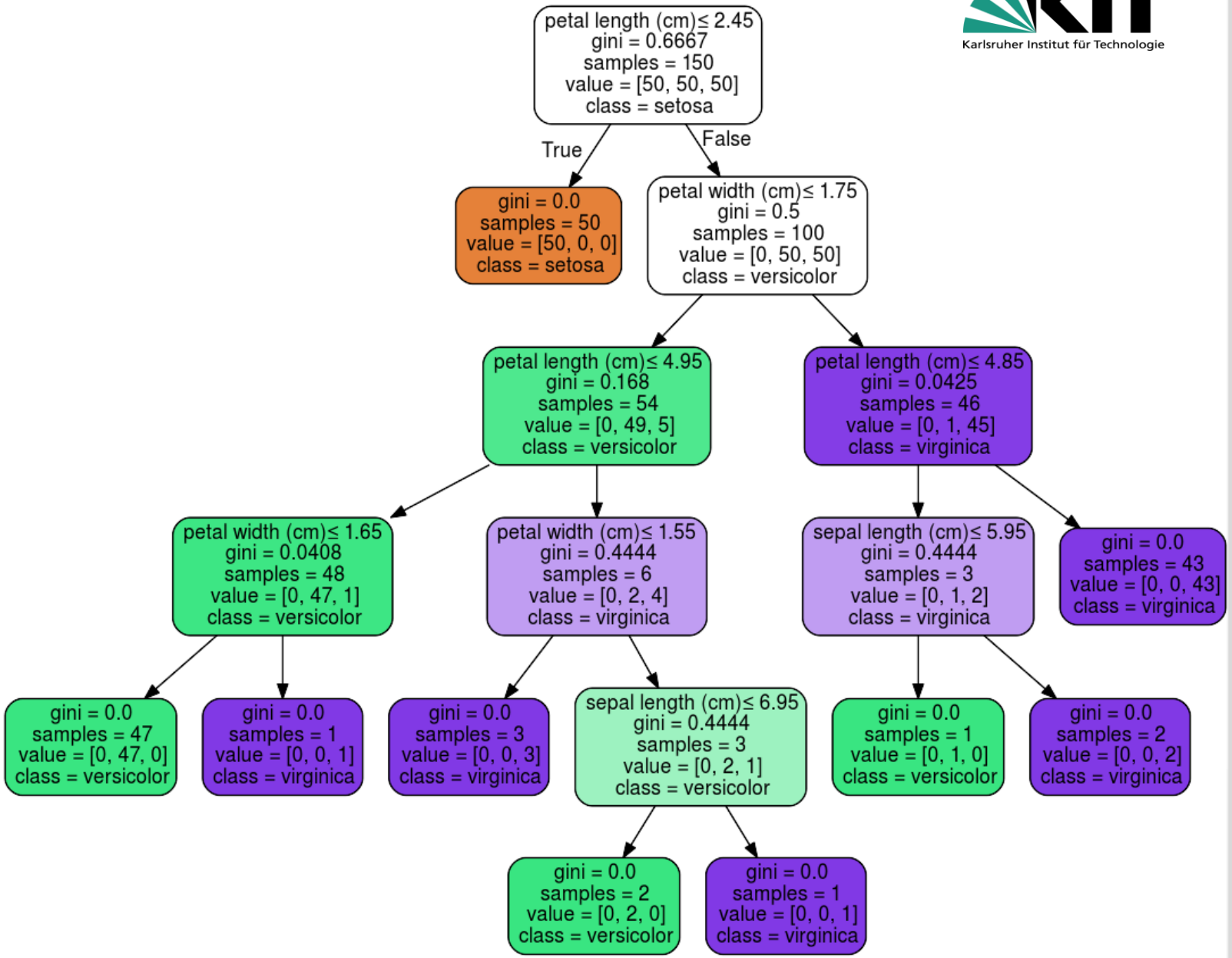
language:	0.4
tweets:	0.0
PhD:	0.950977

Klassifikation

Decision Tree – Iris Datensatz







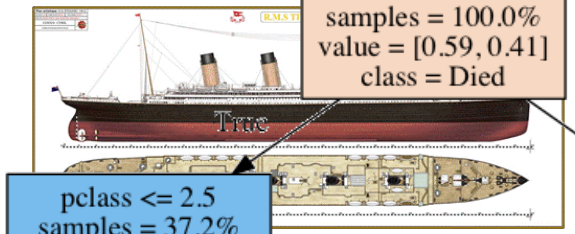
Klassifikation

Decision Tree – Titanic Datensatz

CRISP-DM im Detail: Data Understanding

Merkmale (Features)

- Beispiel: Titanic-Datensatz
  - 11 Features (Name, Alter, Geschlecht, Ticketklasse, Überlebt,…)
    - Survival
    - Pclass
    - Name
    - Sex
    - Age
    - SibSp
    - Parch
    - Ticket
    - Fare
    - Cabin
    - Embarked



75 SoSe 2019 Prof. Dr.-Ing. Eric Sax – Informationstechnik II und Automatisierungstechnik Institut für Technik der Informationsverarbeitung (ITIV)