

DKFZ Heidelberg

Maschinelles Lernen, Klassifikation und diagnostische Tests (6)

Falk Schubert
Intelligente Bioinformatiksysteme
DKFZ Heidelberg

f.schubert@dkfz.de

DKFZ Heidelberg

Kodierung von qualitativen Merkmalen

Umkodierung von qualitativen Merkmalen

- Nominal
- Binäre Kodierung
- Binäre Kodierung
- Aufzählung
- Häufigkeitskodierung
- (MDS)

Qualitative Merkmale

- Merkmal: Antikörperfärbung
- Merkmalsausprägungen:
 - Moderate
 - Strong
 - Neg
 - Weak

	Marker X
Case 1	Moderate
Case 2	Neg
Case 3	Strong
Case 4	Weak

Binäre Kodierung

- Eine neue Spalte für jede Merkmalsausprägung
- Beispiel:

	X_ neg	X_ weak	X_ mod	X_ strong
Case 1	0	0	1	0
Case 2	1	0	0	0
Case 3	0	0	0	1
Case 4	0	1	0	0

DKFZ Heidelberg

5

Falk Schubert

Verbesserte binäre Kodierung

- Führe k-1 neue Spalten für k mögliche Merkmalsausprägungen ein
- Hier:
Neg = 000
Weak = 100
Mod = 110
Strong = 111

	X_ 1	X_ 2	X_ 3
Case 1	1	1	0
Case 2	0	0	0
Case 3	1	1	1
Case 4	1	0	0

DKFZ Heidelberg

6

Falk Schubert

Aufzählung

- Neg=1, Weak=2, Moderate=3, Strong =4
- Problem: Abstände zwischen den Merkmalsausprägungen entsprechen nicht unbedingt der Realität (neg=2*weak)

	Marker X
Case 1	3
Case 2	1
Case 3	4
Case 4	2

DKFZ Heidelberg

7

Falk Schubert

Häufigkeitskodierung

- Kodierung in Abhängigkeit von der Häufigkeit des Auftretens in Bezug auf eine Zielklasse
- Neg=0.05, Weak=0.15, Moderate=0.7, Strong =0.1
where $P(\text{Neg}|\text{Cancer})=0.05$,
 $P(\text{Weak}|\text{Cancer})=0.15$
- Problem:
 - Abstände zwischen den Merkmalsausprägungen

	Marker X
Case 1	0.7
Case 2	0.05
Case 3	0.1
Case 4	0.15

DKFZ Heidelberg

8

Falk Schubert

Zusammenfassung

- Kodierung von qualitativen Merkmalen notwendig für Klassifikatoren wie SVM, Neuronale Netze
 - Binäre Kodierung
 - Häufigkeitskodierung
- Entscheidungsbäume können auch qualitative Merkmale verarbeiten

Softwarerepräsentation

- R
- Clementine