Theorie:

* CRISPDM
  + Phasen und Beschreibung (und Arbeitsschritte)

Praxis:

* Lineare Regression:
  + Grundformel
  + Methode der kleinsten Quadrate
  + Evaluation der Modelle
    - RMSE
    - Korrelation von Attributen -> Einfluss auf bestes Modell
    - Zusammenhang Modellkomplexität <-> RMSE (Overfitting z.B.)
* Entscheidungsbäume
  + Warheits-/Konfussionsmatrix
  + Kostenmatrix

Wahrheitsmatrix

|  |  |  |
| --- | --- | --- |
|  | True Positive | True Negative |
| Pred. Positiv | True Pos. | False Pos. |
| Pred. Negativ | False Neg. | True Negativ |

Präzision Positive: TP/(TP+FP)

Recall Positive: TP/(TP+FN) (Sensitivität)

Recall Negative: TN/(TN+FP) (Spezifität)

Bedingte Wahrscheinlichkeiten: P(H|D)=P(D|H)\*P(H)/P(D) (Wahrscheinlichkeit, dass H wenn D)

P(D)=P(D|H)\*P(H)+P(D|!H)\*P(!H)

Baseline-Modell: Modell ohne Intelligenz als Referenzwert (z.B. alle als Positive einstufen)

Clustern nach Länge & Breite nach k-NN (NN=Next Neighbors)

Aufgabe: gegeben ist Koordinatensystem mit Messpunkten, Vorhersage der Klasse eines Clusters mit k Datenpunkten, ausgehend von n Ausgangspunkten.

Suche nach den k nächstgelegenen Datenpunkten, zählen der Klassenzugehörigkeit, dann Klasseneinteilung des Clusters nach am stärksten vertretener Klasse. Datenpunkte können zu mehreren Clustern gehören.

# Bayesscher Klassifikator/Naive Bayes

Beispiel für 2 Klassen A und B:  
P(A|Daten) > P(B|Daten) dann A, sonst B

|  |  |  |  |
| --- | --- | --- | --- |
| Alter | Einkommen | Akademiker | Kredit gewähren? |
| Jung | Hoch | 0 | 0 |
| Jung | Hoch | 0 | 0 |
| Mittel | Hoch | 0 | 1 |
| Alt | Mittel | 0 | 1 |
| Alt | Niedrig | 1 | 1 |
| Alt | Niedrig | 1 | 0 |
| mittel | Niedrig | 1 | 1 |
| Jung | Mittel | 0 | 0 |
| Jung | Niedrig | 1 | 1 |
| alt | mittel | 1 | 1 |

**Datensatz: Alter=jung**

P(ja|jung)=P(jung|ja)\*P(ja)/P(jung) > P(nein|jung) ?  
//P(jung) muss nicht berechnet werden, da auch in Vergleichsformel enthalten

P(ja)=6/10

P(jung|ja)=1/6 (6 DS sind „ja“, 1 davon ist „jung“)

P(nein)=4/10

P(jung|nein)=3/4

P(ja|jung) = 1/10 (durch P(jung))  
P(nein|jung) = 3/10 (durch P(jung))  
=> Junge Leute haben eine höhere Wahrscheinlichheit, keinen Kredit zu erhalten

Konfidenz C

C(ja)=P(ja|Daten)/(P(Ja|Daten)+P(Nein|Daten)) = 0,1/(0,1+0,3)=1/4

Conf(nein)=P(nein|Daten)/(P(ja|Daten)+P(nein|Daten)) = ¾

**Datensatz: Alter=jung; Einkommen=hoch; Akademiker=nein**

P(ja|{jung, hoch, nein})=P(jung|ja)\*P(hoch|ja)\*P(keinAkademiker|ja)\*P(ja) =1/6\*1/6\*2/6\*6/10=0,555  
P(nein|{jung, hoch, nein})=P(jung|nein)\*P(hoch|nein)\*P(keinAkademiker|nein)\*P(nein) =3/4\*2/4\*3/4\*4/10=11,25  
Confidenz(nein)=95%  
Confidenz(ja)=5%

=>Keine Kredite für junge Menschen mit hohem Einkommen die keine Akademiker sind.

**Datensatz: D={mittel; hoch; Akademiker}**

P(nein|D)=P(mittel|nein)\*P(hoch|nein)…=0/4\*…=0 -> **Zero Frequency Problem** -> Fall tritt in Lerndaten nicht auf.

Lösung: Laplace-Korrektur  
„Hinzufügen“ des Falls. Außerdem hinzufügen des Gegenfalls, um Einfluss niedrig zu halten.  
-> Einfügen: 1 alter, 1 junger, 1 mittlerer mit nein, nicht in Tabelle schreiben, sondern einfach in die Häufigkeiten aufrechnen. => P(mittel|nein)=0+1/4+1 P(jung|nein)= 3+1/4+1 P(alt|nein)=1+1/4+1

**Weiteres Problem: Numerische Attribute (Zusatz! Nicht Prüfungsrelevant)**

Zusätzliches Attribut „Gehalt“

P(5000|ja) -> Annahme einer gausschen Normalverteilung über die Gehälter in der jeweiligen Klasse mit Mittelwert und Standardabweichung

*Übungsaufgaben: Übungsaufgaben Data Mining 4*