Information Mining - Beispiel-Prüfungsfragen

Norbert Fuhr

(neue Fragen sind *kursiv* gedruckt)

- Einführende Fragen: Erzählen sie 3-5 Minuten etwas zu folgendem Thema.
  - Was sind die wesentlichen Problemstellungen des Data Mining (Klassifikation, Assoziationsregeln, Numerische Vorhersagen, Clustering)?
  - Eingabe-Aufbereitung für DM?
  - Wissensdarstellung für die Ausgabe beim DM?
  - Gewinnung von Entscheidungsbäumen
  - Konstruktion von Regeln
  - Erzeugen von Assoziationsregeln
  - Instanzbasiertes Lernen
  - Nummerische Vorhersage
  - Clustering
  - Evaluierung von DM-Verfahren
  - *Data Warehouse*
- Evaluierung:
  - Wie kann von der Erfolgsrate auf einer Teststichprobe auf die Qualität bei zukünftigen Anwendungen schließen (Vertrauensintervall)?
  - Welche Methoden gibt es, um eine begrenzter Datenmenge optimal für Trainings- und Teststichprobe zu nutzen? (Kreuzvalidierung, leave one out, bootstrap)
  - Wie kann DM-Verfahren bzgl. ihrer Qualität vergleichen?
  - Welche anderen Qualitätskriterien für Klassifikationsverfahren gibt es, und wie kann man diese messen? (Vorhersage von Wahrsch., Kosten)
- Entscheidungsbäume:
  - Wie behandelt man numerische Attribute?
  - ... fehlende Werte?
  - Wie funktioniert Pruning?

  - Wie kann man die Fehlerrate abschätzen?
  - Wie kann man einen Baum in Regeln überführen? Welche Probleme treten dabei auf?

- Klassifikationsregeln
  - Kriterien für die Auswahl von Auswertungen?
  - Fehlende Werte und nummerische Attribute?
  - Erzeugung „guter“ Regeln und Entscheidungslisten
  - Wahrscheinlichkeitswert zur Regelevaluation
  - Pruning von Regeln

- Support-Vektor-Maschinen
  - maximal diskriminierende Hyperebene: geometische / mathematische Definition
  - Erweiterung auf nichtlineare Klassengrenzen

- Instanzbasiertes Lernen
  - Wie kann man die Menge der gespeicherten Trainingsinstanzen reduzieren?
  - Wahl einer geeigneten Distanzmetrik?

- Nummerische Vorhersage:
  - Regressionsbäume vs. Modellbäume
  - Welches Kriterium verwendet man zum Aufbau des Modellbaums?
  - Welches zum Pruning?
  - Glättung bei Modellbäumen
  - Wie funktioniert lokal gewichtete lineare Regression?

- *Aufbereitung von Input und Output*
  - *Methoden zur Attribut-Selektion?*
  - *Methoden zur Attribut-Diskretisierung?*
  - *Methoden zur Datentransformation?*
  - *Meta-Lern-Methoden?*
  - *Methoden zur Benutzung unklassifizierter Daten?*

- Clustering
  - Wie funktioniert k-means-Clustering?
  - *Kritischen Punkte bei k-Means sind die Auswahl der Startinstanzen und die Wahl von k — wie kann man diese Probleme angehen?*

- *Welche Bewertungsmaße gibt es für flaches Clustering?*
- Wie funktioniert inkrementelles Clustering?
- Wie funktioniert probabilistisches Clustering?
- *Welche Verfahren gibt es für hierarchisches Clustering, und in welchen Eigenschaften unterscheiden sich diese?*
- *Welche Methoden gibt es für Cluster-Labeling?*
- *Was versteht man unter dichte-basiertem Clustering, und welche Methoden gibt es hierzu?*

Hinweis: Es ist nicht notwendig, dass sie Formeln auswendig wissen. Sie sollten aber die zugrundeliegenden Ideen jeweils wiedergeben können.