Information Mining - Beispiel-Prüfungsfragen

Norbert Fuhr

Fettgedruckte Fragen sollte man beantworten können, um die Prüfung zumindest zu bestehen.

- Einführende Fragen: Erzählen sie 2-3 Minuten etwas zu folgendem Thema.
  - **Was sind die wesentlichen Problemstellungen des Data Mining** (Klassifikation, Assoziationsregeln, Numerische Vorhersagen, Clustering)? Nennen Sie jeweils Anwendungsbeispiele!
  - 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
  - **Welche Klassifikationsverfahren kennen Sie?** Welche davon liefern besodners gute Klassifikationsergebnisse?
  - Data Warehouse
  - Big Data
  - Process Mining
- 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 man DM-Verfahren bzgl. ihrer Qualität vergleichen?** Wie kann man mittels t-Test herausfinden, ob der Unterschied signifikant ist?

  - 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?
  - Methoden zur Benutzung unklassifizierter Daten?
  - Wie kann man Mehrklassenprobleme mittels binärer Klassifikatoren lösen?

- Ensemble-Lernen
  - Wie funktionieren Bagging, Boosting und Metalernen?
  - Was versteht man unter Bias-Varianz-Dekomposition?
  - An welchen zwei Stellen kann man Randomisierung einsetzen?

- Big Data
  - Wie kann man Big Data charakterisieren? (volume, variety, velocity / veracity, variability, venue, vocabulary)
  - Key enablers für Big Data? (inc. of storage capacity, inc. of processing power, availability of data)
  - Warum prozessiert man Big Data häufig in der Cloud?
  - Erläutern Sie die Grundidee von Map-Reduce!
  - Was sind die besonderen Anforderungen an die Data-Mining-Verfahren, wenn man mit Big Data arbeitet?

- Clustering
  - **Wie funktioniert k-means-Clustering?**
  - Kritischer Punkte bei k-Means ist die Auswahl der Startinstanzen wie kann man dieses Probleme angehen?
  - Wie funktioniert probabilistisches Clustering?

- Sequential Pattern Mining
  - Erläutern Sie die Problemstellung und Anwendungsbereiche
  - Was versteht man unter einer Sequenz, einer Subsequenz?
  - Erläutern Sie die apriori-Eigenschaft
  - Erklären Sie die Grundidee eines der vorgestellten Algorithmen

- Process MIning
  - Was ist die Grundidee des Process Mining?
  - Welche Aufgaben gehören zum Process Mining? (basic performance metrics, process model, organizational model, social network, performance characteristics)
  - Erläutern Sie die Grundideen des alpha algorithm (direct succession, causality, parallel, choice / Abb auf Petri-Netze)

- Graph Mining
  - Anwendungsbereiche
  - Erläutern Sie die wichtigsten Konzepte (Knoten, Kanten, Labels, connected subgraph, induced subgraph, graph isomorphism)
  - Aufgabenstellungen: häufige Subgraphen exakt/inexakt in einem/mehreren Graphen, Lernen von Substrukturen mit positiven und negativen Beispielen
  - Was versteht man unter canonical labeling, und wozu kann man es einsetzen?
  - Erklären Sie die Grundidee eines der vorgestellten Algorithmen

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