MS Final Exam Data Mining Study Guide

Text:

1. "Data Mining: Concepts and Techniques", by Jiawei Han and Micheline Kamber 2nd Edition, Morgan Kaufmann Publishers, August 2000.
2. "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", by Ian Witten and Eibe Frank, 2nd edition, Morgan Kaufmann Publishing, 2005.

Material

Witten's book: Chapters 1, 2, 3, 4.1 – 4.4, 4.7 – 4.9, 5
Han's book: Chapters 1-5

Categories:

1. Basic terms, concepts
2. Understand the basic data mining algorithms such as OneR, statistical modeling, ID3. Naïve Bayes, Apriori and Prism, and be able to illustrate them on given data sets (example: Apply the PRISM algorithm on a sample data set to create a classification rule for a particular class.)
3. Know the various ways for evaluating learning algorithms, such as: Holdout estimation, Repeated holdout method, Cross-validation (CV), LOO-CV and stratification, 0.632 bootstrap, Significance tests, Lift charts, and ROC curve
4. Data Cube
   a. Understand N-dimensional data cube and be able to represent a data cube in 2D relational DB or vise versa.
   b. Data cube design using one of the 3 schemas
   c. Specify the basic data cube operations (roll up, slice, etc) that will allow the use to get the desired data

Study Question and Algorithms

Han1

1. The 7 main steps in knowledge discovery
2. The 2 general categories of data mining tasks, based on the kind of patterns to be found: descriptive and predictive (briefly)
3. The 4 schemes for integrating Data Mining System with a DB or DW.
4. Each of the specific patterns that can be mined, such as: concept description, association, classification, cluster analysis, outliner analysis, trend & evolution analysis. (briefly)

Han2

1. The reason for data preprocessing and its major tasks.
2. Missing data: what caused them and how handle them
3. Methods for handling noisy data: binning method, clustering, etc
4. Main tasks in data transformation
5. The characteristics of numeric data (means, midrange, quartile, etc, see hw4)

Han 3

1. Briefly discuss the 4 main features of DW
2. Compare and contrast OLTP and OLAP
3. Know how is a multidimensional point in a data cube space defined
4. Describe and give example of the 3 main DW modeling schema,
5. Describe the three-tier data warehouse architecture
6. Describe and give example of concept hierarchy
7. Describe the different ways of implementing a OLAP server

8. Describe the three scheme for data cube motorization,
9. Specify the basic data cube operations (roll up, slice, etc) that will allow the user to get the desired data (or data aggregation)

Han 4

1. Name and briefly describe the 4 general techniques for cube computation
2. Briefly describe the general method of multiway array aggregation
3. Illustrate the general idea of AOI and apply to a specific example

Witten1-3

1. The 4 basic style of learning
2. Give example of various forms of mined knowledge: Decision table, decision tree, classification rules, association rules, regression tree, IBL, clusters,
3. How to derive classification rules from decision trees (give example)
4. The general forms of rules with exceptions

Witten 4

1. Illustraste the basic data mining algorithms such as: OneR, Statistical Modeling, ID3
1. Discuss the basic concept of a covering algorithm
2. Apply Prism algorithm on sample data to create a good rule (similar to problem 1 of hw5)
3. Discuss the concepts of frequent item, frequent item set, support, confidence
4. Apply the Aprori algorithm on sample data (similar to problem 2 of hw5)
5. Discuss the k-NN algorithm
6. Discuss the k-means algorithm

Witten 5

Be able to describe/discuss

1. how the lift chart works
2. how the ROC chart works
3. ways to measure the info retrieval