Data and Web Mining Test 1 - 2019 - Memo

Statistics 1A (Namibia University of Science and Technology)

NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY
Faculty of Computing and Informatics
Department of Computer Science

13 Storch Street
Private Bag 13388
Windhoek
NAMIBIA

T: +264 61 207 2258
F: +264 61 207 9258
E: dcs@nust.na
W: www.nust.na

Data and Web Mining (DWM710S) – Test 1 Memorandum

Duration: 1 Hour
Marks: 50

Date: 22 March 2019

Questions

1. Illustration with a practical example, please explain, "Data mining turns a large collection of data into knowledge". [2 Marks]

   Example: A search engine (e.g. Google) receives hundreds of millions of queries every day. Google's Flu Trends uses specific search terms as indicators of flu activity.

2. Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers. [2 Marks]

   There are many outlier detection methods. We propose two methods for fraudulence detection:

   a) Statistical methods (also known as model-based methods): Assume that the normal transaction data follow somestatistical (stochastic) model, and then data not following the model are outliers.

   b) Clustering-based methods: Assume that the normaldata objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.

3. What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)? [6 Marks]

   One challenge to data mining regarding performance issues is the efficiency and scalability of data mining algorithms. Data mining algorithms must be efficient and scalable in order to effectively extract information from large amounts of data in databases within predictable and acceptable running times. Another challenge is the parallel, distributed, and incremental processing of data mining algorithms. The need for parallel and distributed data mining algorithms has been brought about by the huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods. Due to the high cost of some data mining processes, incremental data mining algorithms incorporate database updates without the need to mine the entire data again from scratch.

4. Discuss by providing an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis? [6 Marks]

A department store, for example, can use data mining to assist with its target marketing mail campaign. Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products. Data query processing is used for data or information retrieval and does not have the means for finding association rules. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

5. Patterns can be presented in the form of association rules. Please explain the following association rule: [4 Marks]

computer => antivirus_software [support = 2%, confidence = 60%]

The information that customers who purchase computers also tend to buy antivirus at the same time.

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules.

A support of 2% means that 2% of all the transactions under analysis show that computer and antivirus software are purchases together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

6. Imagine that you are a sales manager at Technology Firm, and you are talking to a customer who recently bought a PC and a digital camera from the store. What should you recommend to her next? [3 Marks]

Information about which products are frequently purchased by your customers following their purchases of a PC and a digital camera in sequence would be very helpful in making your recommendation

7. In the context of understanding data, clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes.

With the following themes, please demonstrate clusters with a practical example.

a) Climate [3 Marks]

Understanding the Earth's climate requires finding patterns in the atmosphere and ocean. To that end, cluster analysis has been applied to find patterns in the atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate.

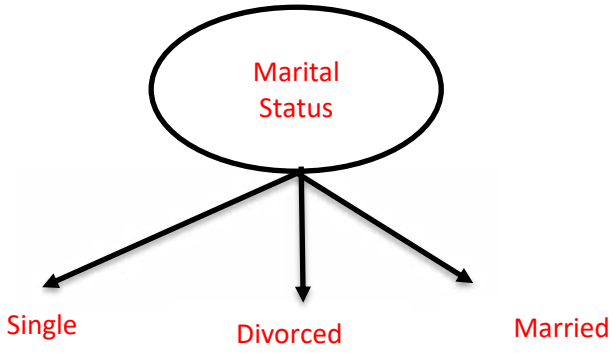
b) Business [3 Marks]

Businesses collect large amounts of information on current and potential customers. Clustering can be used to segment customers into a small number of groups for additional analysis and marketing activities.

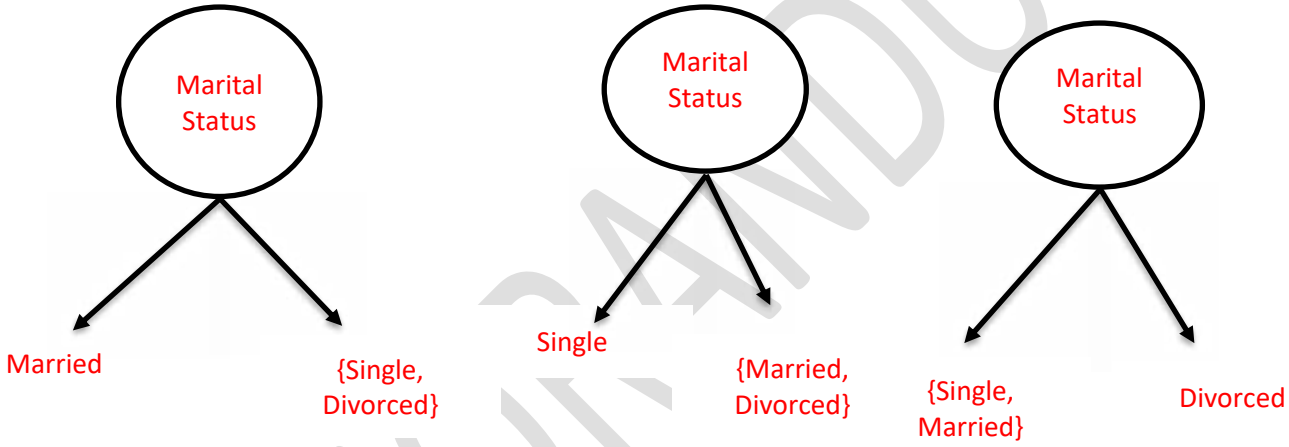
8. Decision tree induction algorithms must provide a method for expressing an attribute test condition and its corresponding outcomes for different attribute types.

By using decision trees, please demonstrate each of the below attributes:

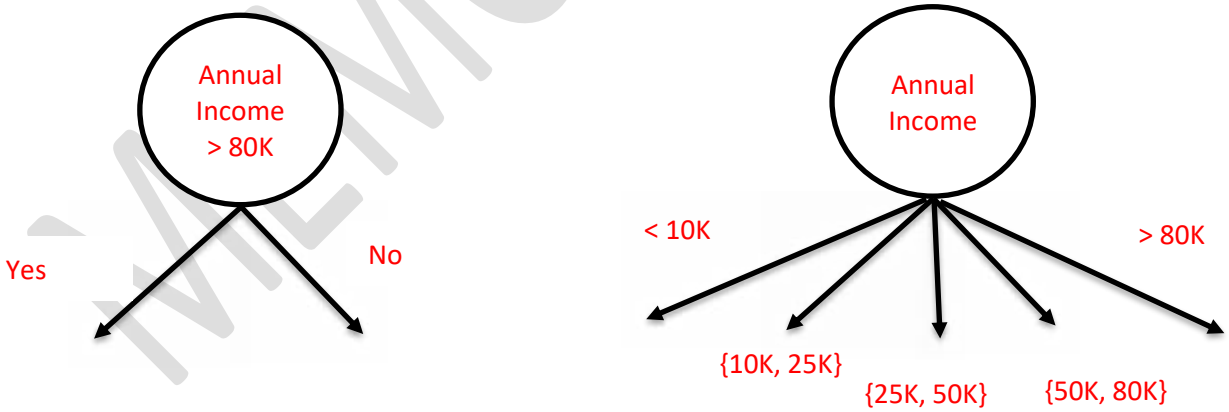
a) Test condition for Nominal Attributes – (the multiway split) [4 Marks]



b) Test condition for Nominal Attributes – (the binary split) [4 Marks]



c) Test condition for Continuous Attributes [4 Marks]



9. Suppose your task as a Software Engineer at Big-University is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and their cumulative grade point average (GPA).

Describe the architecture you would choose. What is the purpose of each component of this architecture? [6 Marks]

A data mining architecture that can be used for this application would consist of the following major components:

• A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.

• A database or data warehouse server which fetches the relevant data based on users' data mining requests.

• A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.

• A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.

• A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.

• A graphical user interface that allows the user an interactive approach to the data mining system.

End of Paper