Yuan Wang
440 Data Mining
Prof. Jiebo Luo
Homework 1

Part 1

1.1
A:
Data mining is not a hype, it is a methodology to get knowledge from massive data in information age.
B:
Data mining is an interdisciplinary subject, which requires multiple technologies from other domains such as databases, statistics, machine learning, and pattern recognition.
C:

Yes, it's also the result of the evolution of machine learning research. Based on the process of Machine learning and statistics, the data mining evolution is: data input, data pre-processing (including data integration, normalization, feature selection, dimension reduction), data mining (including pattern discovery, association & correlation, classification, clustering, outlier analysis), post-processing (pattern evaluation, pattern selection, pattern interpretation, pattern visualization).[1]
D:
A process of knowledge discovery includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation.[2]

1.2
Database and Data warehouse are both relational data system, they both consist of interrelated data and a set of software program to manage and access the data.
However, database and data warehouse have many differences. Data warehouse stores data by multidimensional data structure, called data cube. Data warehouse also allows different level of abstraction, which supports drill down and roll up operations. Database stores data by a collection of tables, each of table consists a set of attributes and usually stores a large set of tuples.[3]

1.4
Data mining can be crucial for google's online advertisement. For example, Google conducts online advertisement for University of Rochester. It is very important for google to classify the customers and their specific demands such as degree, major, tuition fee, etc. Some of the information can be directly collect by google from U of R. However, the customer information is hard to extract directly from customers. Google needs to generate that information by data mining, such as customers' search history, related online comments, interests, etc.

[1] *Ch 1.2, Data Mining: Concepts and Techniques", 3/E, by Jiawei Han, Michelin Kamber and Jian Pei.*
[2] *Ch 1.3, Data Mining: Concepts and Techniques", 3/E, by Jiawei Han, Michelin Kamber and Jian Pei.*
[3] Ibid

1.5

Discrimination & Classification

Both data discrimination and classification use the similar methods to collect data.
Data Classification summarize the general characteristics or features of a target class of data.
Data Discrimination compares the target class with one or multiple contrasting classes.[4]

Characterization & Clustering

Characterization is a summarization of the general characteristics or features of a target class of data.
Clustering analyzes data objects without consulting class labels, in many cases, class-labeled data may not exist at the beginning.[5]

Classification & Regression

Classification & Regression both analyze class labeled data set.
Classification predicts categorical labels, whereas regression predicts missing or unavailable numerical prediction and class label prediction.[6]

1.7

There are many methods to detect outliers, such as statistical measures, distance measures, and density-based methods.[7]
In fraudulent detection, the statistical measure seems more reliable. For example, the credit card company could use the statistical data to define the consumption behavior for different group of customers. This consumption behavior is relatively stable for an identical group of customers, the abnormal expenditure could trigger fraudulent detection for further analysis.

1.9

Comparing a small amount of data to a huge amount of data, there are two critical challenges needed to consider. First, data mining algorithms must be efficient and scalable to extract information from huge amounts of data which may get from various resources. Second, the huge amount of data brings in the tremendous data size, the wide distribution of data, and the complexity of mining methods, which requires the development of parallel and distributed data-intensive mining algorithms.[8]

***Note:** This homework referred knowledge from "Data Mining: Concepts and Techniques", 3/E, by Jiawei Han, Michelin Kamber and Jian Pei."*

[4] *Ch 1.4, Data Mining: Concepts and Techniques", 3/E, by Jiawei Han, Michelin Kamber and Jian Pei*
[5] Ibid
[6] Ibid
[7] Ibid
[8] *Ch 1.7.3, Data Mining: Concepts and Techniques", 3/E, by Jiawei Han, Michelin Kamber and Jian Pei*