Data Warehousing: Theory and Practice

Exam 2

Instructor Dr. Yong Shi Spring 2021

ISQA 8700 (Remote Learning) Due on April 27, 2021

*This is an open-book exam. Do and show all possible steps for the following problems. Hand in your test to Mr. Haotian Li at UNO Canvas or* [haotianli@unomaha.edu](mailto:haotianli@unomaha.edu)*:*

*This is an open-book exam about Chapter 6-8 and 13 (four pages). Do and show all possible steps for the following problems:*

# Part I Essay on (50 points)

What do you think about data mining theory and applications (two pages)?

The main categories involved in data mining are :

1. Theoretical Foundations of Data Mining:

* **Data Reduction** − The basic idea of this theory is to reduce the data representation which trades accuracy for speed in response to the need to obtain quick approximate answers to queries on very large databases. Some of the data reduction techniques are as follows −
  + Singular value Decomposition
  + Wavelets
  + Regression
  + Log-linear models
  + Histograms
  + Clustering
  + Sampling
  + Construction of Index Trees
* **Data Compression** − The basic idea of this theory is to compress the given data by encoding in terms of the following −
  + Bits
  + Association Rules
  + Decision Trees
  + Clusters
* **Pattern Discovery** − The basic idea of this theory is to discover patterns occurring in a database. Following are the areas that contribute to this theory −
  + Machine Learning
  + Neural Network
  + Association Mining
  + Sequential Pattern Matching
  + Clustering
* **Probability Theory** − This theory is based on statistical theory. The basic idea behind this theory is to discover joint probability distributions of random variables.
* **Probability Theory** − According to this theory, data mining finds the patterns that are interesting only to the extent that they can be used in the decision-making process of some enterprise.
* **Microeconomic View** − As per this theory, a database schema consists of data and patterns that are stored in a database. Therefore, data mining is the task of performing induction on databases.
* **Inductive databases** − Apart from the database-oriented techniques, there are statistical techniques available for data analysis. These techniques can be applied to scientific data and data from economic and social sciences as well.

Statistical Data Mining:

* **Regression** − Regression methods are used to predict the value of the response variable from one or more predictor variables where the variables are numeric. Listed below are the forms of Regression −
  + Linear
  + Multiple
  + Weighted
  + Polynomial
  + Nonparametric
  + Robust
* **Generalized Linear Models** − Generalized Linear Model includes −
  + Logistic Regression
  + Poisson Regression

The model's generalization allows a categorical response variable to be related to a set of predictor variables in a manner similar to the modelling of numeric response variable using linear regression.

* **Analysis of Variance** − This technique analyzes −
  + Experimental data for two or more populations described by a numeric response variable.
  + One or more categorical variables (factors).
* **Mixed-effect Models** − These models are used for analyzing grouped data. These models describe the relationship between a response variable and some co-variates in the data grouped according to one or more factors.
* **Factor Analysis** − Factor analysis is used to predict a categorical response variable. This method assumes that independent variables follow a multivariate normal distribution.
* **Time Series Analysis** − Following are the methods for analyzing time-series data −
  + Auto-regression Methods.
  + Univariate ARIMA (AutoRegressive Integrated Moving Average) Modeling.
  + Long-memory time-series modeling.

Visual Data Mining:

* Data Visualization
* Data Mining

Visual data mining is closely related to the following −

* Computer Graphics
* Multimedia Systems
* Human Computer Interaction
* Pattern Recognition
* High-performance Computing

Generally, data visualization and data mining can be integrated in the following ways −

* **Data Visualization** − The data in a database or a data warehouse can be viewed in several visual forms that are listed below −
  + Boxplots
  + 3-D Cubes
  + Data distribution charts
  + Curves
  + Surfaces
  + Link graphs etc.

The major large applications of Data Mining are:

1. **Future Healthcare:**

* Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

1. **Market Basket Analysis:**

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer’s needs and change the store’s layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done

1. **Education:**

* There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students’ future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

1. CRM:

Customer Relationship Management is all about acquiring and retaining customers, also improving customers’ loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results

# Part II Essay on (25 points)

What do you understand Decision Tree Algorithm (one page)?

Decision Tree is used to build classification and regression models. It is used to create data models that will predict class labels or values for the decision-making process. The models are built from the training dataset fed to the system.

Information Gain

Gini Index

**1. Information Gain**

When we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.

**Entropy**  
Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.

**Classification and Regression Tree(CART):**

It is a dynamic learning algorithm which can produce a regression tree as well as a classification tree depending upon the dependent variable.

**Classification Analysis:**

Data Classification is a form of analysis which builds a model that describes important class variables. For example, a model built to categorize bank loan applications as safe or risky. Classification methods are used in machine learning, and pattern recognition.

Application of classification includes fraud detection, medical diagnosis, target marketing, etc. The output of the classification problem is taken as “Mode” of all observed values of the terminal node.

**Regression Analysis :**

Regression analysis is used for the prediction of numeric attributes.

Numeric attributes are also called continuous values. A model built to predict the continuous values instead of class labels is called the regression model. The output of regression analysis is the “Mean” of all observed values of the node.

**A two-step process is followed, to build a classification model.**

1. In the first step i.e. learning: A classification model based on training data is built.
2. In the second step i.e. Classification, the accuracy of the model is checked and then the model is used to classify new data. The class labels presented here are in the form of discrete values such as “yes” or “no”, “safe” or “risky”.

**Working of decision tree:**

A decision tree is a supervised learning algorithm that works for both discrete and continuous variables. It splits the dataset into subsets on the basis of the most significant attribute in the dataset. How the decision tree identifies this attribute and how this splitting is done is decided by the algorithms.

The most significant predictor is designated as the root node, splitting is done to form sub-nodes called decision nodes, and the nodes which do not split further are terminal or leaf nodes.

In the decision tree, the dataset is divided into homogeneous and non-overlapping regions. It follows a top-down approach as the top region presents all the observations at a single place which splits into two or more branches that further split. This approach is also called a greedy approach as it only considers the current node between the worked on without focusing on the future nodes.

The decision tree algorithms will continue running until a stop criteria such as the minimum number of observations etc. is reached.

Once a decision tree is built, many nodes may represent outliers or noisy data. Tree pruning method is applied to remove unwanted data. This, in turn, improves the accuracy of the classification model.

# Part III Essay on (25 points)

Describe how data mining will affect your life in the future (one page).

#### **Data Mining**

Data mining is an extraction of interesting (potentially useful) or knowledge from the massive amount of data.

The wide availability of vast amounts of data and the imminent need for turning such data into useful information and knowledge.

Data mining is defined as a process used to extract usable data from a larger set of any raw data which implies analysing data patterns in large batches of data using one or more software.

#### **Real life Examples in Data Mining that would affect anyone’s life and necessity of implication of data mining:**

Following are the various real-life examples of data mining,

**1. Shopping Market Analysis**

There is a huge amount of data in the shopping market, and the user needs to manage large data using different patterns. Market basket analysis is a modelling technique is used to do the analysis. Market basket analysis is a modelling technique based on a theory that if you buy a group of items, you are more likely to buy another group of things. This technique may allow the retailer to understand the purchase behaviour of a buyer. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

**2. Stock Market Analysis**

There is a vast amount of data to be analysed in the stock market. So, data mining technique is used to model those data to do the analysis.

**3. Weather forecasting analysis**

Weather forecasting system uses an enormous amount of historical data for prediction. As there is a processing of enormous amount data, one must have to use the suitable data mining technique.

**4. Fraud Detection**

Due to the size of the data, traditional methods of fraud detection are time-consuming and complicated. Data mining helps in providing meaningful patterns and turning data into information.

**5. Intrusion Detection**

Data mining can help to improve intrusion detection by adding a level of focus to anomaly detection. It supports an analyst to distinguish activity from common everyday network activity.

**6. Financial Banking**

A tremendous amount of data is supposed to be generated with new transactions in computerised banking. Data mining can donate to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information.

**7. Surveillance**

Video surveillance is used in a day to day life almost at every place for security perception. Data mining is used in video surveillance as we need to deal with a large amount of collected data.

**8. Online Shopping**

In online shopping, E-commerce companies like Amazon, Flipkart, Snapdeal, Myntra, and many more uses Data Mining and Business Intelligence to offer cross-sells and up-sells through their websites, who use sophisticated mining techniques to drive there, ‘People who viewed that product, also liked this’ functionality. Data mining is used to identify customers loyalty by analysing the data of customer’s purchasing activities such as the data of frequency of purchase in a period, a total monetary value of all investments and when was the last purchase.

**9. Criminal Investigation**

Criminal Investigation is a process that intentions to identify crime characteristics. Crime analysis includes discovering and detecting crimes and their relationships with criminals. The large volume of crime datasets and the complexity of relationships between them have made criminology a suitable field for applying data mining techniques.

**10. Bioinformatics**

Data Mining approaches are well suited for Bioinformatics, as it contains a massive amount of data. The mining of biological data aids to extract useful knowledge from massive datasets gathered in biology, and other related life sciences areas such as medicine and neuroscience.

**11. Health Care and Insurance**

The growth of the insurance industry entirely depends on the ability to convert data into the knowledge, information or intelligence about customers, competitors, and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully.