**Data Mining – CSC240**

**Name – Hizkia Febianto**

Question 1.1

Data mining is a process of extracting knowledge or insights from a vast amount of data. Analogous with mining gold from rocks and sand, the process of data mining will result in small nuggets of valuable information from a great deal of raw data.

Data mining is not a hype. The field emerges from the necessity to handle big amounts of data being collected in the digital network. Nowadays, information about a person, from personal information to consumption habit, is being monitored and recorded by many companies. The number of data being generated from this data collection is enormous, making it impossible to understand the knowledge behind the data without help from a computer system. Data mining provides a set of tools and methods of processing data and generate useful information from the data.

Since data mining is interdisciplinary, its development is also a collective result of the development of other disciplines, such as machine learning and statistics. Same as data mining, machine learning and statistics have evolved to accommodate high amount of data. Machine learning saw rapid growth in 1990s, due to the invention of World-Wide-Web and large data gathered on the internet. From that point in time, machine learning required more automation and computing power. Many new algorithms have also been proposed to increase results.

The first step of data mining is data collection process. Data are collected from various sources and stored in multiple databases. Before being stored in the database, the data have to be cleaned from any noise and inconsistency. This process is called *data cleaning*. Cleaned data are then integrated and stored in the databases. This is the *data integration* process. At this point, the data are still distributed in many databases so they need to be gathered into a single place to make the analysis easier. Relevant data are retrieved from the databases and consolidated into forms appropriate for data mining. The data are stored in a single location under a unified schema. This process consists of two steps, namely *data selection* and *data transformation*. The unified schema in which relevant data are stored is called data warehouse. After the data are gathered in the data warehouse, the *data mining* process is applied to extract patterns in the data. After that, the patterns are *evaluated* and knowledge from the pattern is *presented* and visualized.

Question 1.2

Both database and data warehouse store and manage interrelated data based on their attributes. Both of them use SQL to query the data. Their main difference is in how the data are being organized. In a database, the data are stored in complex tables. Each table has data attributes as columns and data records as rows and it is possible that there is a relationship between two tables. Since tables in the database can have different numbers of columns and different numbers of rows, a database is too complex to be used directly in the analysis. Data warehouse integrates data from multiple databases and stores them in a unified schema. Since the data are organized in the same structure, accessing the data become faster and the analysis can be done more efficiently. In a data warehouse, the data are typically organized in a multidimensional data structure, called a data cube. This structure shows the relationships between data attributes in its axes. A data cube provides a multidimensional views of data attributes and sometimes is precomputed, resulting in fast access of the data.

Question 1.4

Suppose *AllElectronics* wants to give discount coupons to its customers. To maximize profits from this initiatives, they have to send the coupons to customers who are likely to use the coupon and have a high chance of spending a lot of money in their purchases. Mistakenly targeted coupons will incur costs for the company since they have to pay for the postal fees. For this purpose, *AllElectronics* looks at the historical data about how their customers react when they are given a coupon. The data contain a list of customers who used their coupon and the amount of money that they spent. The data also tell the company which customers did not use their coupon. From this data, *AllElectronics* can build a model to predict whether a selected customer with a set of characteristics will use a coupon and spend a lot of money.

This kind of pattern is very difficult to generate simply by using data query. We can query a list of customers that respond to the coupon well but we cannot extract patterns in the characteristics of those customers. Furthermore, data query cannot predict the likelihood of a customer responding to a coupon given a set of characteristics of the person.

Question 1.5

Both *discrimination and classification* consider two mutually-exclusive classes/categories and a set of features that generally characterizes each class. The difference between discrimination and classification is that discrimination focuses on understanding features of each contrasting class while classification assigns a class to a data record depending on its features.

Both *characterization* *and clustering* find similarities in the features that characterize a particular class or category. The difference between these two processes is on whether the class is initially defined or not. In characterization process, the class is defined in the beginning. The algorithm then analyzes prominent features that characterize the class. In the clustering process, the classes do not have to be defined in the beginning. The algorithm will analyze the features and classify the data into several classes based on feature similarity. Since it does not require classes to be defined in the beginning, the clustering process is an example of unsupervised learning algorithm.

Both *classification and regression* uses a model to predict the class label of a data record. The model is build using a set of training data. The difference between these two processes is that classification predicts discrete classes while regression predicts a continuous value. Regression typically produces numerical data values as outputs while regression gives a finite number of discrete classes.

Question 1.7

The first method that can be used to detect fraudulent transactions is classification. Each transaction is classified as either valid transaction or fraudulent transaction. Using a training data set, a model is developed to classify transaction into one of the classes. The problem with classification is that the number of fraudulent transactions is much lower than the number of valid transaction. This class imbalance will reduce the prediction accuracy of the model. To solve this problem, methods for handling imbalanced classes may be used.

The second method that can be used is clustering method. In this method, we assume that valid transactions belong to large and dense clusters, whereas outliers belong to small clusters or do not belong to any clusters. This method is better than classification method since it is not affected by class imbalance. Since the principle of clustering method is to maximize the intra-class similarity and to minimize the interclass similarity, this method is good for detecting outlier classes.

Question 1.9

The first challenge of mining a huge amount of data is computing power requirement. Many data mining techniques involve multiple iterations and extensive calculations. Computing power does not seem like a problem when we use a small amount of data. A few hundred records might be analyzed within seconds with a particular data mining technique. However, the same technique might need several days when we analyze data in the order of billions.

The second challenge is data storage and accessibility. A big amount of data requires a big storage space. The highest capacity of hard disk with current technology is still in the order of terabyte. So, if we want to store data with a volume of petabyte or higher, we need to use hundreds or thousands of hard disks. This brings other challenges: how to control all of these hard disks and how to find a record among thousands of hard disks. Access time will also become a challenge since finding a record in this distributed system may require a significant amount of time.

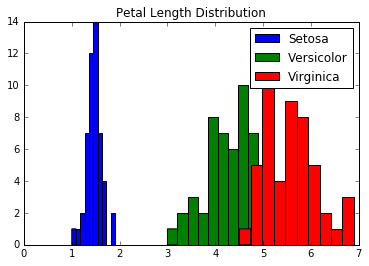
**Exploratory Data Analysis**

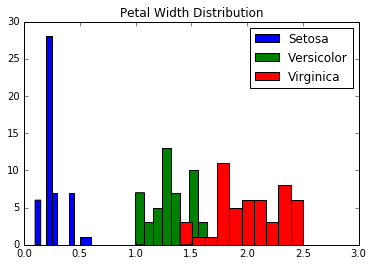
Iris dataset[[1]](#footnote-1) has 150 records of iris flower from three classes. Each record has 5 attributes, namely petal length, petal width, sepal width, sepal length, and class. In this analysis, we will calculate mean and standard deviation of each attribute, plot the distribution of each attribute, and find out the best attribute that can be used to classify iris flower.

Summary statistics of each type of iris is shown in the table below.

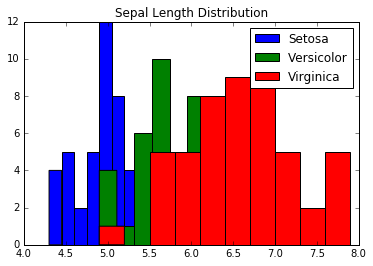
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Statistics** | **Petal length** | **Petal width** | **Sepal length** | **Sepal width** |
| Iris Setosa | Mean | 1.46 | 0.24 | 5.00 | 3.42 |
| Stdev | 0.17 | 0.11 | 0.35 | 0.38 |
| Iris Versicolor | Mean | 4.26 | 1.33 | 5.94 | 2.77 |
| Stdev | 0.47 | 0.20 | 0.52 | 0.31 |
| Iris Virginica | Mean | 5.55 | 2.03 | 6.59 | 2.97 |
| Stdev | 0.55 | 0.27 | 0.64 | 0.32 |

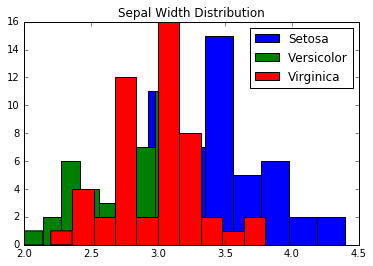
Iris Setosa has the smallest petal size among other types of iris. We can see on the table that petal length and width are pretty distinct. Later analysis will prove that petal size is the best attribute to classify iris flowers. Sepal width and sepal length have very close mean values. Therefore, it is difficult to distinguish iris classes based on sepal size.



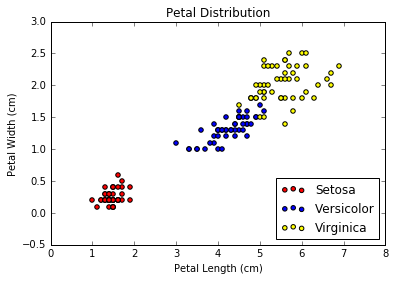


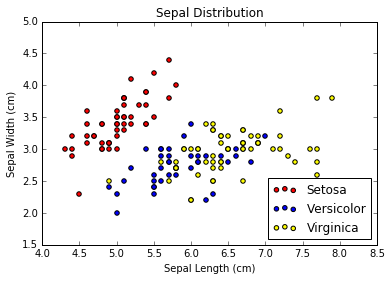
The distributions of petal length and petal width are shown in the graphs above. They are pretty distinguishable and only overlap a little bit. Iris Setosa is very different than the others. It has petal length values in the range of 1 – 2 cm and petal width values in the range of 0.1 – 0.5 cm. Iris Versicolor and Iris Virginica are also easy to distinguish based on petal length and width. If we combine petal length and petal width to create a scatter plot, the classification of the classes become even clearer.





From two graphs above, we can see the distribution of sepal width and sepal length. Unlike petal, sepal does not provide a good distinction between each class because the distributions highly-overlap.





As we can see from the graphs above, petal length and petal width separate iris classes almost perfectly. We can create a good classification model from petal size and get an accurate classification of iris classes. On the contrary, sepal size does not provide a clear separation. We can classify Iris Setosa well, but we cannot classify Iris Versicolor and Iris Virginica with a high accuracy using sepal dimensions.

1. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science [↑](#footnote-ref-1)