# Assignment 1:

- Q1. Give two examples, apart from those given in the slides, for each of the following:

a) Data mining from the commercial viewpoint

- It is clear that we are using data mining on many commercial fields for many different purposes. For example, in business and economic environment, we apply data mining techniques when we want to do any sales campaigns, so if we found that the results are as our expectation, then we can start the campaigns, otherwise the decision will be to cancel it.[1]

Also, we use advanced techniques of data mining in strategic management of enterprises. Some decision cannot be taken by top management without applying some analysis based on data stored on the company's database or other data from external sources to ensure taking the right decision.[1]

[1]:Andronie, Mihai and Crisan, Daniel, (2010), Commercially Available Data Mining Tools used in the Economic Environment, Database Systems Journal, 1, issue 2, p. 45-54.

b) Data mining from the scientific viewpoint

- A good example is to use data mining in a biological science such as in clinical trials, to detect the conclusion that relates to the effectiveness of a specific drug that used as a treatment of a disease. Also, to find any patterns related to diseases occurrences.[2]

[2]: Kamath, C. (2009). Scientific data mining: A practical perspective (p. 18). Philadelphia: Society for Industrial and Applied Mathematics.

-Another example is using data mining in earth sciences since these data may be satellite images or observations. We apply data mining for some purposes such as land cover classification for monitoring changes, earthquakes detection, storm warning, or cloud detection.[3]

[3]: Grossman, R., Kamath, C., Kegelmeyer, P., Kumar, V., & Namburu, R. (2013). Data mining for scientific and engineering applications (Illustrated ed., Vol. 2, p. 10). Springer Science & Business Media.

==========

- Q2. Differentiate between classification of data and clustering of data with the help of suitable examples.

- **Classification** is a predictive method, in which we have a set of records which can be called (training set), and in every record we have attributes but one of them will be the class. Then we need to use the other attributes in a function to find a model for our class, which means predicting the attribute value. For example: a company needs to analyze its employees history to know who will renew his subscription in internet service. So, a model will be constructed to predict the categorical labels which are Yes or No.[4]

[4]: Data Mining Classification & Prediction. (n.d.). Retrieved September 19, 2015, from http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

- **Clustering** is a descriptive method, in which we have a set of data or records with a set of attributes which have a similarity or dissimilarity among them. So, we divide the set of data or the records into group of similar objects or records in a cluster, and dissimilar ones in another cluster. Then we give each group or cluster a specific label. For example, if you have an educational institute, you can make groups of your students based on their subscription patterns. So, when you want to offer specific prices or discounts for specific students, the clustering process will help you to choose the suitable group for this offer.[5]

[5]: Data Mining Cluster Analysis. (n.d.). Retrieved September 19, 2015, from http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

==========

❖ Q3. Why do we need preprocessing of the data? Explain any 4 data preprocessing techniques.

- Data preprocessing is an important step in the whole data mining process because it prepares initial dataset to go through mining process smoothly and effectively. The data at the beginning may be in low quality such as noise, inconsistency, or missing values, and that will affect the result of data mining and the decisions upon it. So, preprocessing will ensure that the data quality is high according to (accuracy, completeness, consistency, timeliness, believability, and interpretability), which will facilitate the mining process and improve its efficiency to get high quality results.[6]

- **Data preprocessing techniques:**

**Data Cleaning:** It is a technique used to clean data by handling missing values, removing or smoothing noisy data, removing outliers, and solving the inconsistencies problems in order to prepare data for mining process.

**Data Integration:** It is another technique used to merge and integrate data from multiple data sources (databases, data cubes, or files) into a coherent store. In this technique we need to use some statistical functions to remove any redundancies and to detect any inconsistencies.

**Data Reduction:** This technique is used to save time during data analysis process. It is to obtain a reduced representation of the data set, but in smaller volume, and this representation will produce analytical results that same or almost same as the original dataset. Data reduction strategies include dimensionality reduction, in which data encoding schemes are applied to get a reduced representation. Another strategy is numerosity reduction, in which data are replaced by alternative, smaller representations through parametric or nonparametric models.

**Data Transformation and data discretization:** data is transformed or combined into forms that suitable for mining. So, it needs to perform some operations such as normalization to scale attribute data to fall within small range, smoothing and removing the noise from data, aggregation and summarization operations, and discretization by dividing the range of a continuous attribute into intervals.[6]

[6]: Han, J., & Kamber, M. (2012). Data mining concepts and techniques, third edition (3rd ed., pp. 84-116). Waltham, Mass.: Morgan Kaufmann.

==========

❖ Q4. Explain in detail the 5 number summary of distribution (i.e. Minimum, Q1, Median, Q3, Maximum) of a box plot.

- To create a box plot we need to specify five number summary of distribution which are:

**Minimum:** the smallest individual observation or the smallest element on the data which forms the lower extreme point.

**Q1:** first quartile, it forms the $25^{th}$ percentile which divides the lowest 25% of data.

**Median:** second quartile, it forms $50^{th}$ percentile which is the middle or center of data.

**Q3:** third quartile, it forms the $75^{th}$ percentile which divides the lowest 75% of data.

**Maximum:** the largest individual observation or the largest element of data that forms the upper extreme point.

So, the Q1 and Q3 forms the ends of the box, and the median marked by a line within it. Then we make two lines (whiskers) but outside the box extend to the minimum and maximum points.[7]

[7]: Han, J., & Kamber, M. (2012). Data mining concepts and techniques, third edition (3rd ed., pp. 48-49). Waltham, Mass.: Morgan Kaufmann.

==========

❖ Q5. Give any two situations in which a distribution of data is negatively skewed in one and positively skewed in the other. You can think of any real life example.

- When we look to the comparison of number of children who can start walking, we will find that the age from 12 to 18 months have the most frequency, which form the right side of the distribution. However, there are some children can start walking at the age of 9 to 11 months, and that means towards the left side of the distribution, which form the long tail to the left. So, the **distribution is negatively skewed**.

- In a case I have a car rent shop, and we made a distribution to see how many cars that cost more or less money among different cars, we will notice that we have thirty cars cost between 100SR/day and 150SR/day. However, there are two cars that cost 250SR/day, and one car which cost 300SR/day. So, most of high number of cars are concentrated on the 100 and 150. Which means that the bulk of distribution is towards the left side of the distribution which are the lower cost. On the other hand, there will be less number of cars on the higher cost (250 and 300), and that will form a long tail towards the right, which means that this **distribution is positively skewed**.

==========