1. What is data mining? In your answer, address the following:
   (a) Is it another hype?
   (b) Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?
   (c) We have presented a view that data mining is the result of the evolution of database technology. Do you think that data mining is also the result of the evolution of machine learning research? Can you present such views based on the historical progress of this discipline? Do the same for the fields of statistics and pattern recognition.
   (d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Answer: Data mining refers to the process or method that extracts or "mines" interesting knowledge or patterns from large amounts of data.

(a) Is it another hype? Data mining is not another hype. Instead, the need for data mining has arisen due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, data mining can be viewed as the result of the natural evolution of information technology.

(b) Is it a simple transformation of technology developed from databases, statistics, and machine learning? No. Data mining is more than a simple transformation of technology developed from databases, statistics, and machine learning. Instead, data mining involves an integration, rather than a simple transformation, of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis.

(c) Explain how the evolution of database technology led to data mining. Database technology began with the development of data collection and database creation mechanisms that led to the development of effective mechanisms for data management including data storage and retrieval, and query and transaction processing. The large number of database systems offering query and transaction processing eventually and naturally led to the need for data analysis and understanding. Hence, data mining began its development out of this necessity.

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery. The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

• Data cleaning, a process that removes or transforms noise and inconsistent data
• Data integration, where multiple data sources may be combined
• Data selection, where data relevant to the analysis task are retrieved from the database
• Data transformation, where data are transformed or consolidated into forms appropriate for mining
• Data mining, an essential process where intelligent and efficient methods are applied in order to extract patterns
• Pattern evaluation, a process that identifies the truly interesting patterns representing knowledge based on some interestingness measures
• Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user

2. Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.

Answer: **Characterization** is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University first year computing science students, which may include such information as a high GPA and large number of courses taken.

Discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

Association is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

$$major(X, \text{computing science}) \Rightarrow owns(X, \text{“}personal\ computer\text{”})$$

$$[support = 12\%, confidence = 98\%]$$

where X is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence, or certainty) that a student in this group owns a personal

computer. Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold. Additional analysis can be performed to uncover interesting statistical correlations between associated attribute-value pairs.

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. It predicts categorical (discrete, unordered) labels.

Regression, unlike classification, is a process to model continuous-valued functions. It is used to predict missing or unavailable numerical data values rather than (discrete) class labels.

Clustering analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Outlier analysis is the analysis of outliers, which are objects that do not comply with the general behavior or model of the data. Examples include fraud detection based on a large dataset of credit card transactions.

3. 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 and discuss which one is more reliable.

Answer:

There are many outlier detection methods. More details can be found in Chapter 12. Here we propose two methods for fraudulence detection:

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

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

It is hard to say which one is more reliable. The effectiveness of statistical methods highly depends on whether the assumptions made for the statistical model hold true for the given data. And the effectiveness of clustering methods highly depends on which clustering method we choose.

4. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows. Compute an approximate median value for the data. Show your solution.

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Answer:

$L_1 = 20,\ n = 3194,\ (\sum_f)_l = 950,\ freq\_median = 1500,\ \text{width} = 30,\ \text{median} = 30.94\ \text{years}.$

5. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   (a) What is the mean of the data? What is the median?
   (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
   (c) What is the midrange of the data?
   (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
   (e) Give the five-number summary of the data.
   (f) How is a quantile-quantile plot different from a quantile plot?

Answer:

a. The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 809/27 = 30.$ The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: **25**.
b. This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are **25** and **35**.
c. The midrange (average of the largest and smallest values in the data set) of the data is: (70 + 13)/2 = 41.5
d. The first quartile (corresponding to the 25th percentile) of the data is: **20**. The third quartile (corresponding to the 75th percentile) of the data is: **35**.
e. The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: **13, 20, 25, 35, 70**.

f. A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.
A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line (y = x) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

6. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median and standard deviation of age and %fat.
(b) Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

Answer:

a) For the variable age the mean is **46.44**, the median is **51**, and the standard deviation is **12.85**. For the variable %fat the mean is **28.78**, the median is **30.7**, and the standard deviation is **8.99**.

b)

age	23	23	27	27	39	41	47	49	50
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.15	-0.18	0.27
age	52	54	54	56	57	58	58	60	61
z-age	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.77

The correlation coefficient is 0.82. The variables are positively correlated.

7. Using the data for age given in item no. 5, answer the following:
   (a) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].
   (b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
   (c) Use normalization by decimal scaling to transform the value 35 for age.

Answer:

a) Using the corresponding equation with minA = 13, maxA = 70, new minA = 0,new maxA = 1.0, then v = 35 is transformed to v' = 0.39.

b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. Using the corresponding equation where A = 809/27 = 29.96 and σA = 12.94, then v = 35 is transformed to v' = 0.39.

c) Using the corresponding equation where j = 2, v = 35 is transformed to v' = 0.35.