KDD HW 01 - KDD Assignment 1 Ssolutions

Knowledge Disc In Databases (University of North Carolina at Charlotte)

Homework 1

What is data mining? In your answer, address the following:

Data mining refers to the process of extracting or mining interesting knowledge or patterns from large amounts of data.

(a) **Is it another hype?**

No, Data mining is not another hype. “We are living in the information age” is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining.

(b) Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?

No. Data mining is not a simple transformation of technology developed from databases, statistics, and machine learning. Instead, it involves an integration of data 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, and information retrieval and so on.

(c) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Steps involved in Data mining when viewed as Knowledge Discovery process.

- **Data cleaning**- a process that removes or transforms noise and inconsistent data
- **Data integration**- where data from heterogeneous data sources is combined for mining purpose.
- **Data selection**- where data relevant to the analysis task are retrieved from the database
- **Data transformation** - where data is transformed or consolidated into forms suitable 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. Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

A departmental store, can use data mining to assist with its target marketing mail campaign. Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products. Data query processing is used for data or information retrieval and does not have the means for finding association rules. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

3. Based on your observation, describe another possible kind of knowledge that needs to be discovered by data mining methods but has not been listed in this chapter. Does it require a mining methodology that is quite different from those outlined in this chapter?

For example, one may propose partial periodicity as a new kind of knowledge, where a pattern is partial periodic if and if only some offsets of a certain time period in a time series demonstrate some repeating behavior.

We can use Sentimental Analysis to predict whether the citizens of a country are Happy or Sad. We can search all tweets in a twitter using a key word Happy or Sad. We can arrive at a decision as to why the person is happy or sad. These method uses text mining technique to get the data set. After analyzing the data set, various government organization can have a record of what is happening in a particular part of the country and take necessary actions to resolve it if majority of citizens are sad about the same issue.

4. It is important to define or select similarity measures in data analysis. However, there is no commonly- accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless,

seemingly different similarity measures may be equivalent after some transformation.

Suppose we have the following two-dimensional data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(a) Consider the data as two-dimensional data points. Given a new data point, x = (1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

Euclidean distance

Distance = ((1.5-1.4) ^2 – (1.7-1.6)^ 2) ^0.5 = 0.1414

Manhattan (city block, L_1 norm) distance

Distance = (1.5-1.4) + (1.7-1.6) = 0.2

Supremum distance

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Cosine similarity

cos(d_1, d_2) = ($d_1 \bullet d_2$) / $||d_1||$ $||d_2||$

cos(d1,d2)=((1.5*1.4)+(1.7*1.6))/(((1.5)^2 +(1.7)^2)*((1.4)^2+(1.6)^2)) =0.99999

	Euclidean	**Manhattan**	**Supreme**	**Cosine Sim**
X1	0.1414	0.2	0.1	0.99999
X2	0.6708	0.9	0.6	0.99575
X3	0.2828	0.4	0.2	0.99997
X4	0.2236	0.3	0.2	0.99903
X5	0.6083	0.7	0.6	0.96536

These values produce the following rankings of the data points based on similarity:

Euclidean distance: x_1, x_4, x_3, x_5, x_2

Manhattan distance: x_1, x_4, x_3, x_5, x_2

Supremum distance: x_1, x_4, x_3, x_5, x_2

Cosine similarity: x_1, x_3, x_4, x_2, x_5

(b) Use three ways (min-max, z-score, and decimal scaling) to normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

Min and Max values from the data set are 1.0 and 2.0.

New min =0 and Max=1.

Using Min-max normalization: to [new_min_A, new_max_A]

The data set in normalized to below points.

	A1	A2
x1	0.5	0.7
x2	1.0	0.9
x3	0.6	0.18
x4	0.2	0.5
x5	0.5	0.0

The query point is normalized as (0.4, 0.6)
Euclidean distances for the normalized data set are below.

	Euclidean Dist.
x1	0.1414
x2	0.6708
x3	0.2828
x4	0.2236
x5	0.6083

Order of Euclidean distances for the normalized(Min Max) data set is: x_1, x_4, x_3, x_5, and x2 which is same as the order before normalization.

Decimal scaling normalization

Where *j* is the smallest integer such that makes |v'| < 1

Let j=1

	A1	A2
x1	0.15	0.17
x2	0.2	0.19
x3	0.16	0.18
x4	0.12	0.15
x5	0.15	0.1

The query point is normalized as (0.14, 0.16)
Euclidean distances for the normalized data set are below.

	Euclidean Dist.
x1	0.0141
x2	0.067
x3	0.0282
x4	0.0223
x5	0.0608

Order of Euclidean distances for the normalized(Decimal Scaling) data set is: x_1, x_4, x_3, x_5, and x2 which is same as the order before normalization.

Z score Normalization (μ: mean, σ: standard deviation):

Mean: 1.57

SD: 0.3056

	A1	A2
x1	-0.2290	0.425
x2	1.407	1.079

x3	0.098	0.752
x4	-1.21	-0.229
x5	-0.229	-1.86

The query point is normalized as (-0.5562, 0.098)
Euclidean distances for the normalized data set are below.

	Euclidean Dist.
x1	0.4625
x2	2.194
x3	0.9249
x4	0.7309
x5	1.986

Order of Euclidean distances for the normalized (Z norm) data set is: x_1, x_4, x_3, x_5, and x2 which is same as the order before normalization.