**Data Mining**

**Home Assignment – 1**

Name : Sri Satya Krishna Atluri

OU ID : 113474131

Course ID : CS 5593

**Problem 1: For each of the following activities, answer if it is or if it is not a data mining task, and justify your answer.**

1. Find out how a stock price will change based on the stock’s past performance.
2. Find out the frequency of amino acids types of human Hemoglobin protein.
3. Identify all students who are likely to get an “A” in the Operating Systems course.
4. Recommend movies to customers based on their movie viewing pattern.
5. Identify all customers who have made more than 5 purchases at a local store.

**Solution:**

1. Yes, finding how a stock price changes in the future based on the stock’s past performance is a Data Mining task. This is a data mining task since it involves the process of analyzing the past information of the stock and then predicting or generating new information about the stock such as how its price varies. The raw data from the previous performances is converted into useful information for the future predictions. Hence this is a data mining task.
2. Yes, finding frequency of amino acids types of human Hemoglobin protein is a Data Mining task. The frequency of amino acids in Hemoglobin varies from person to person and it is not a constant value as it keeps changing based on the condition of the human. This task requires the process of finding patterns in the changes of frequency of amino acids of a period of time to predict the frequency of amino acids. Hence this is a data mining task.
3. Yes, identifying students who are likely to get an “A” in operating systems course is a Data Mining Task. This task involves prediction of grades for operating systems course. These predictions can be done by analyzing the details of students who got an “A” in the previous exams and various other factors. Hence this is a data mining task.
4. Yes, recommending movies to customers based on their movie viewing pattern is a Data Mining task. This task is similar to the functioning of Netflix and Amazon networks where the history of the user (such as what the user have viewed in the past helps in knowing the user interests and giving suggestions based on them). If a user has viewed mostly action movies in the past, it is easier to know that the user is more likely to watch action movies than other genre movies. Since this is task involving finding patterns in the user browsing history and predicting or suggesting a movie, it is a data mining task.
5. No, identifying the customers who made more than 5 purchases at a local store is not a Data Mining task. This task is just about searching. And retrieval of past data and it can be done using a database command. This is not a data mining task since it does not involve generating new data from existing data.

**Problem 2: For each attribute given, classify its type as:**

* **discrete or continuous AND**
* **qualitative or quantitative AND**
* **nominal, ordinal, interval, or ratio**

1. Cell phone brands
2. IQ levels
3. The states of the United States
4. The prices of laptops
5. The result of whether a person has passed a Driving Exam. The result can only be “Pass” or “Fail”.

**Solution:**

1. Discrete, Qualitative, Nominal.
2. Discrete, Qualitative, Ordinal.
3. Discrete, Qualitative, Nominal.
4. Continuous, Quantitative, Ratio.
5. Discrete, Qualitative, Nominal.

**Problem 3: Perform the following tasks:**

1. Calculate the similarity/distance measure between the two vectors given below using each of the five methods: Simple Matching Coefficient, Jaccard Coefficient, Cosine, Correlation, and Hamming distance. Show your work in detail.

x = (1,0,1,1,0,1,0), y = (1,1,0,1,0,0,1)

1. Assume that the following set of two-dimensional points are randomly sampled from the same multivariate Gaussian distribution: {(-2.05, 2.32), (-0.41, 5.36), (0.72, 3.62), (0.15, 3.1), (-1.3, 4.1), (-3.7, 2.8)}. Compute the Euclidean distance and the Mahalanobis distance between the first two points in the set. Show your work in detail.
2. The bank (iBank) wants to analyze the results of a survey that it has conducted to study the customers’ experiences with its services. The survey has 60 questions, each of which has five possible answers (Very Dissatisfied; Somewhat Dissatisfied; Neutral; Somewhat Satisfied; Very Satisfied).
   1. How would you convert this data into a form suitable for association analysis? In particular, what types of attributes would you have and how many are there in total?
   2. Answer the same question in part (a), but with the assumption that both non-zero and zero values are equally important.

**Solution:**

1. Given two vectors are:

x = (1,0,1,1,0,1,0), y = (1,1,0,1,0,0,1)

Computing the similarity/distance measure using the below five methods,

**Simple Matching Coefficient:**

SMC = number of matches / number of attributes

SMC = (F11 + F00) / (F01 + F10 + F11 + F00)

where, F00 – number of attributes where x is 0 and y is 0

F11 – number of attributes where x is 1 and y is 1

F01 – number of attributes where x is 0 and y is 1

F10 – number of attributes where x is 1 and y is 0

Here, F00 = 1, F11 = 2, F01 = 2, F10 = 2

Therefore,

SMC = (2 + 1) / (2 + 2 + 2 + 1)

SMC = 3 / 7

SMC = 0.428

**Jaccard Coefficient:**

J = number of 11 matches / number of non-zero attributes

J = (F11) / (F01 + F10 + F11)

where, F11 – number of attributes where x is 1 and y is 1

F01 – number of attributes where x is 0 and y is 1

F10 – number of attributes where x is 1 and y is 0

Here, F11 = 2, F01 = 2, F10 = 2

Therefore,

J = (2) / (2 + 2 + 2)

J = 2 / 6

J = 0.333

**Cosine:**

cos(d1, d2) = (d1 **.** d2) / ||d1|| ||d2||

where, d1 and d2 are vectors, ||d|| is length of vector d and **.** indicated dot product.

Here,

d1 **.** d2 = 1\*1 + 0\*1 + 1\*0 + 1\*1 + 0\*0 + 1\*0 + 0\*1

d1 **.** d2 = 1 + 0 + 0 + 1 + 0 + 0 + 0 = 2

||d1|| = (1\*1 + 0\*0 + 1\*1 + 1\*1 + 0\*0 + 1\*1 + 0\*0)0.5 = (4)0.5 = 2

||d2|| = (1\*1 + 1\*1 + 0\*0 + 1\*1 + 0\*0 + 0\*0 + 1\*1)0.5 = (4)0.5 = 2

Therefore,

cos(d1, d2) = (2) / (2)(2) = 2 / 4 = 0.5

**Correlation:**

correlation(x, y) = (x’ **.** y’) / (n – 1)

where, xk’ = (xk – mean(x)) / std(x)

yk’ = (yk – mean(y)) / std(y)

Here, mean(x) = 4 / 7 = 0.57, mean(y) = 4 / 7 = 0.57

std(x) = (((1-0.57)2+(-0.57)2+(1-0.57)2+(1-0.57)2+(-0.57)2+(1- 0.57)2+(-0.57)2)/ 7)0.5

std(x) = ((0.184 + 0.324 + 0.184 + 0.184 + 0.324 + 0.184 + 0.324)/7)0.5

std(x) = ((0.736 + 0.972)/7)0.5

std(x) = (1.708/7)0.5

std(x) = (0.244)0.5

std(x) = 0.493

std(y) = (((1-0.57)2+(1-0.57)2+(-0.57)2+(1-0.57)2+(-0.57)2+(- 0.57)2+(1-0.57)2)/ 7)0.5

std(y) = ((0.736 + 0.972)/7)0.5

std(y) = (1.708/7)0.5

std(y) = (0.244)0.5

std(y) = 0.493

Therefore,

correlation(x, y) = ((1-0.57)\*(1-0.57) + (-0.57)\*(1-0.57) + (1-0.57)\*(- 0.57) + (1-0.57)\*(1-0.57) + (-0.57)\*(-0.57) + (1-0.57)\*(-0.57) + (- 0.57)\*(1-0.57)) / (0.493\*0.493) / 6

correlation(x, y) = ((0.184 – 0.245 – 0.245 + 0.184 + 0.324 – 0.245 – 0.245) / 0.243) / 6

correlation(x, y) = ((-0.288)/0.243)/6

correlation(x, y) = (-1.185)/6

correlation(x, y) = -0.1975

**Hamming Distance:**

Hamming distance = F01 + F10

where, F01 – number of attributes where x is 0 and y is 1

F10 – number of attributes where x is 1 and y is 0

Here, F01 = 2, F10 = 2

Therefore,

Hamming distance = 2 + 2

Hamming distance = 4

1. Given gaussian distribution is:

{(-2.05, 2.32), (-0.41, 5.36), (0.72, 3.62), (0.15, 3.1), (-1.3, 4.1), (-3.7, 2.8)}

**Euclidean Distance:**

dist =

dist = ((-2.05+0.41)2 + (2.32-5.36)2)0.5

dist = ((1.64)2+(-3.04)2)0.5

dist = (2.689+9.24)0.5

dist = (11.929)0.5

dist = 3.453

**Mahalanobis Distance:**

mahalanobis(p, q) =

where, C is the covariance matrix, p and q are the vectors.

Cj,k =

Here,

= (-6.59)/6 = -1.098 (mean of first attribute)

= 21.3/6 = 3.55 (mean of second attribute)

C1,1 = (1/5)(-0.9522+0.6882+1.8182+1.2482-0.2022-2.6022) = 2.61

C1,2 = 0.76

C2,1 = 0.76

C2,2 = 1.17

Therefore,

mahalanobis(p, q) = ((-1.64, -3.04))0.5

mahalanobis(p, q) = ((-1.64, -3.04))0.5

mahalanobis(p, q) = ((-1.64, -3.04))0.5

mahalanobis(p, q) = (-0.259 + 8.220)0.5

mahalanobis(p, q) = 2.821

**Problem 4: Using R and the Credit Approval Data Set from the UCI dataset repository website** [**https://archive.ics.uci.edu/ml/datasets/Credit+Approval**](https://archive.ics.uci.edu/ml/datasets/Credit+Approval)**, perform the following tasks:**

1. Write a function that estimates the missing values in the dataset as follows: for each attribute that has one or more missing values, replace every missing value in that attribute with the *mean* value if the attribute is continuous and with the *mode* value if the attribute is categorical. Then use this function to estimate all the missing values in the dataset.
2. Using the six attributes A2, A3, A8, A11, A14, and A15 of the preprocessed data, do the following:
   1. Draw a random sample *with replacement* of size N=100 from the dataset.
   2. Create a matrix of scatter plots showing the correlations between all possible pairs of attributes from the above six attributes. Using the scatter plots, comment on which pairs of attributes exhibit the highest correlation. The axes of your plots must be properly labeled.
   3. Calculate the correlations between all possible pairs formed by those six attributes.
   4. Normalize those six attributes using the Z-score normalization.
   5. Create another matrix of scatter plots showing the correlations between all possible pairs of attributes from the six *normalized* attributes.
   6. Re-calculate the correlations between all possible pairs formed by the six *normalized* attributes.
3. Using your scatter plots of Task 2 as examples, explain *in detail* whether or not the Z-score normalization affects the correlation of the data.