GIFT UNIVERSITY

G U J R A N W A L A

(Chartered by the Govt. of the Punjab, Recognized by HEC)

EXA-115

Department of Computer Science

Data Mining (DS-306A)

Final Term Examination
Fall-2022

Instructor: Dr. Muhammad Faheem

Time: 01:40 Hrs **Total Marks: 50**

Solution

Instructions to Candidates:

- Candidates are required to sit on the seats assigned to them by the invigilators.
- Do not open this question paper until you have been told to do so by the Invigilator.
- Please fill in exam specific details in space provided (both Question Paper and Answer Sheet).
- This is a Closed Book Exam. “Closed book examinations’” refer to examinations where the candidate may not take into the examination room any study materials (including textbooks, study guides, lecture notes, printed notes from web pages, hand written notes and any audio/visual aid).
- **There are 8 questions. Attempt all questions.**
- Do not write anything on question paper except Name and Roll Number.

Paper Setter’s Signature **HoD Signature**

Question 1: **(5 Marks)**

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.

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.

Question 2:

(5 Marks)

Suppose we have the following two-dimensional data set:

	A1	**A2**
x1	1.5	1.7
x2	2	1.9
x3	1.6	1.8
x4	1.2	1.5
x5	1.5	1

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.

Use Equation (2.6) to compute the Euclidean distance, Equation (2.7) to compute the Manhattan distance, Equation (2.8) to compute the supremum distance, and Equation (2.9) to compute the cosine similarity between the input data point and each of the data points in the data set. Doing so yields the following table

	Euclidean Dist.	Manhattan Dist.	Supermum Dist.	Cosine Similarity
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: x1, x4, x3, x5, x2
Manhattan distance: x1, x4, x3, x5, x2
Supremum distance: x1, x4, x3, x5, x2
Cosine similarity: x1, x3, x4, x2, x5

Question 3:

Briefly outline how to compute the dissimilarity between objects described by the following:
(a) Nominal attributes
(b) Asymmetric binary attributes
(c) Numeric attribute
(d) Term-frequency vector

Nominal variables
A nominal variable is a generalization of the binary variable in that it can take on more than two states.
The dissimilarity between two objects i and j can be computed using the simple matching approach as in (2.4):

$$d(i, j) = p - m / p, \quad (2.4)$$

where m is the number of matches (i.e., the number of variables for which i and j are in the same state), and p is the total number of variables. Alternatively, we can use a large number of binary variables by creating a new binary variable for each of the M nominal states. For an object with a given state value, the binary variable representing that state is set to 1, while the remaining binary variables are set to 0.

Asymmetric binary variables
If all binary variables have the same weight, we have the contingency Table 2.1. Use

object i \ object j	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Table 10.1: A contingency table for binary variables.

the Jaccard coefficient, where the number of negative matches, t, is considered unimportant and thus is ignored in the computation, that is,

$$d(i, j) = r + s / q + r + s \quad (2.5)$$

Numeric Attribute
Use Euclidean distance or Manhattan distance.

Term-frequency vectors
To measure the distance between complex objects represented by vectors, it is often easier to abandon traditional metric distance computation and introduce a nonmetric similarity function.
For example, the similarity between two vectors, x and y, can be defined as a cosine measure, as follows:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||} \quad (2.9)$$

where $\boldsymbol{x}^t$ is a transposition of vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$,[1] $||\boldsymbol{y}||$ is the Euclidean norm of vector $\boldsymbol{y}$, and s is essentially the cosine of the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

Question 4: **(5 Marks)**

Explain the min-max and z-normalization process using examples.

Refer to Chapter 3 for examples provided in details.

Question 5: **(5 Marks)**

Suppose a group of 12 sales price records has been sorted as follows:
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.
Partition them into three bins by each of the following methods:
(a) equal-frequency (equal-depth) partitioning
(b) equal-width partitioning

(a) equal-frequency (equidepth) partitioning
Partition the data into equidepth bins of depth 4:
Bin 1: 1: 5, 10, 11, 13
Bin 2: 15, 35, 50, 55
Bin 3: 72, 92, 204, 215

(b) equal-width partitioning
Partitioning the data into 3 equi-width bins will require the width to be (215 – 5)/3 = 70. We uget:
Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72
Bin 2: 92
Bin 3: 204, 215

Question 6: **(10 Marks)**

A database has five transactions. Let min sup D 60% and min conf D 80%.

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

Paper Setter's Signature **HoD Signature**

i. For Apriori, one finds the following frequent itemsets, and candidate itemsets (after deletion as a result of has infrequent subset):
L1 = {E, K, M, O, Y}
C2 = {EK, EM, EO, EY, KM, KO, KY, MO, MY, OY}
L2 = {EK, EO, KM, KO, KY}
C3 = {EKO}
L3 = {EKO}
C4 = ∅
L4 = ∅

Finally resulting in the complete set of frequent itemsets:
{E, K, M, O, Y, EK, EO, KM, KO, KY, EKO}

ii. For FP-growth, one finds the following:
Note: CPB(item) denotes the Conditional Pattern Base of that item, and CFPT(item) denotes the Conditional FP-Tree of that item.

L = {{E: 4}, {K: 4}, {M: 3}, {O: 3}, {Y: 3}}
CPB(Y) = {{E,K,M,O: 1}, {E,K,O: 1}, {K,M: 1}}
CFPT(Y) = ⟨K: 3⟩
Generates FPs: {K,Y: 3}

CPB(O) = {{E,K,M: 1}, {E,K: 2}}
CFPT(O) = ⟨E: 3, K: 3⟩
Generates FPs: {E,K,O: 3}, {K,O: 3}, {E,O: 3}

CPB(M) = {{E,K: 2}, {K: 1}}
CFPT(M) = ⟨K: 3⟩
Generates FPs: {K,M: 3}

CPB(K) = {{E: 4}}
CFPT(K) = ⟨E: 4⟩
Generates FPs: {E,K: 4}

Which finally results in the complete set of frequent itemsets:
{ {E: 4}, {K: 4}, {M: 3}, {O: 3}, {Y: 3}, {K,Y: 3}, {E,K,O: 3}, {K,O: 3}, {E,O: 3}, {K,M: 3}, {E,K: 4} }

You should identify that FP-growth is more efficient because it is able to mine in the conditional pattern bases, which may substantially reduce the sizes of the data sets to be searched.
However, when working with small data sets like the one given (especially when working by hand) you may feel like Apriori is more ”efficient.”

Question 7: **(10 Marks)**

Explain the algorithm steps of the decision tree induction and find the following attribute selection measures of the given tables.

a) Info(D), D= (yes, no)
b) Gain (income)
c) Gain (student)
d) Gain (credit rating)

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

Refer to section 8.2.2. The example is provided with the explanation of each step.

Question 8: **(5 Marks)**

What is boosting? State why it may improve the accuracy of decision tree induction.

Boosting is a technique used to help improve classifier accuracy. We are given a set S of s tuples. For iteration t, where t = 1, 2, . . . , T, a training set St is sampled with replacement from S. Assign weights to the tuples within that training set. Create a classifier, Ct from St. After Ct is created, update the weights of the tuples so that the tuples causing classification error will have a a greater probability of being selected for the next classifier constructed. The final boosting classifier combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy. It can be proved that with the number of weak classifiers increasing, the training error of combination boosting classifier drops.

End of Question Paper.

Paper Setter's Signature **HoD Signature**