NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY OF APPLIED SCIENCES

DEPARTMENT OF INFORMATICS AND ANALYTICS

DATA MINING AND DATA WAREHOUSING

SIA1202

Second Semester Examinations 2024

This examination paper consists of 4 pages.

Time Allowed: **3 hours**

Total Marks: **100**

Examiner's Name: **PROFESSOR N. GASELA**

INSTRUCTIONS

INSTRUCTIONS: ANSWER ALL QUESTIONS IN SECTION A AND ALL QUESTIONS SECTION B

MARK ALLOCATION

QUESTION	MARKS
1.	**20**
2.	**20**
3.	**20**
4.	**20**
5.	**20**
Total of 4 Questions	**100**

SECTION A. [Answer all questions: 1/2 Each Total Marks = 20]

Question 1
Name a data mining tool that works with a canvas to design you data mining workflow.
(A) Cortana
(B) KNIME
(C) Python
(D) Weka

Question 2
Which of the following statements is correct?
(A) Leave-One-Out is a method aimed at missing values.
(B) Leave-One-Out is a better choice than 10-fold cross-validation when dealing with large data (large n).
(C) Leave-One-Out is simply cross-validation with k = n.
(D) When I'm not satisfied with the accuracy using 10-fold cross-validation, I can try Leave-One-Out to get a better score.

Question 3
What does the Apriori principle refer to?
(A) The fact that the Apriori algorithm is one of the most principal ones.
(B) The principle that there is no one algorithm that works best on all datasets.
(C) The fact that a superset of an itemset X has at most the support of X.
(D) The notion that the prior of the target is a good lower bound for the accuracy of a classifier.

Question 4
What is the formula for the entropy of a nominal attribute?

(A) $\sum_i p_i lg(p_i)$

(B) $-\sum_i p_i lg(p_i)$

(C) $-\sum_i lg(p_i)$

(D) $plg(p)$

Question 5
Which of the following statements is correct?
(A) The information gain of an attribute can be any positive number.
(B) The information gain of an attribute can be less than 0.
(C) The information gain of an attribute is at most 1.
(D) The information gain of an attribute is bounded (from above) by the entropy of the target.

Story A, Frequent Pattern Mining
The following 'story' asks you to complete an itemset lattice with the following labels: I= infrequent itemset, F=frequent itemset, M=maximal frequent itemset, C=closed frequent itemset. The upcoming questions test whether you computed the labels correctly.
The lattice provided here can be used as a draft, and doesn't need to be handed in. Only the answers to the multiple choice questions are relevant.
Given a transactional database with the following itemsets over fA; : : : ;Eg, and a minimal support minsup = 0:3:

tid	Items
1	$\{A, C, B, E\}$
2	$\{D\}$
3	$\{A, B, E\}$
4	$\{A, C\}$
5	$\{A, D, C\}$
6	$\{C, B, E\}$
7	$\{D, C\}$
8	$\{A, C, B, E\}$
9	$\{B, E\}$
10	$\{D, C\}$

Question 6 (Story A)
Which itemset in Story A is frequent?
(A) $\{D, E\}$
(B) $\{A, D\}$
(C) $\{A, B\}$
(D) $\{B, D\}$

Question 7 (Story A)
What are the maximal (frequent) itemsets in the dataset of Story A?

Question 8 (Story A)
Which of the following statements is **not** correct?
(A) $\{A, C\}$, $\{C, D\}$, $\{A, B, E\}$, and $\{B, C, E\}$ are closed itemset.
(B) $\{A, B\}$, $\{B\}$ and $\{E\}$ are closed itemset.
(C) $\{A\}$, $\{C\}$, $\{D\}$, and $\{B, E\}$ are closed itemset.
(D) $\{A, C, D\}$ is a closed itemset.

Question 9
What is the definition of a maximal (frequent) itemset?
(A) An itemset is maximal frequent if none of its immediate supersets has the same support.
(B) An itemset is maximal frequent if none of its immediate subsets is frequent.
(C) An itemset is maximal frequent if none of its immediate supersets is frequent.
(D) An itemset is maximal frequent if none of its immediate subsets has the same support.

Question 10
Rank the following in order of entropy, from low to high: 1) a person's gender, 2) whether they own a Ferrari, 3) a person's social security number, 4) a person's highest education.

(A) a person's gender, whether they own a Ferrari, a person's social security number, a person's highest education.
(B) whether they own a Ferrari, a person's gender, a person's highest education, a person's social security number.
(C) a person's gender, a person's highest education, whether they own a Ferrari, a person's social security number.
(D) a person's social security number, a person's highest education, a person's gender, whether they own a Ferrari.

Question 11
What is a disadvantage of using histograms to estimate the density of an attribute?

(A) Bin boundaries can be placed at unfortunate locations, causing empty bins, or too full bins.
(B) It performs worse than Kernel Density Estimation.
(C) It assumes a normal distribution.
(D) It is an unsupervised method

Question 12
Which of the methods below is an example of an unsupervised learning algorithm?

(A) k-NN.
(B) k-means Clustering.
(C) Subgroup Discovery.
(D) Linear Regression.

Question 13
Looking at the descriptions, as well as the feature values in the table below, which data types match which features?

_ genre: Contains various names of movie styles.
_ rating: Movies are rated on a 5 point scale from very bad to very good.

_ gross: Money that the movie made.
_ cinema: If the movie was shown in cinemas.

id	genre	rating	gross	cinema
1	horror	very bad	5000	0
2	drama	good	8000	1
3	comedy	very good	9000	1

(A) genre = nominal, rating = ordinal, gross = ordinal, cinema = nominal
(B) genre = nominal, rating = nominal, gross = numeric, cinema = binary
(C) genre = ordinal, rating = nominal, gross = numeric, cinema = binary
(D) genre = nominal, rating = ordinal, gross = numeric, cinema = binary

Question 14
Which of the following statements about clustering is correct?

(A) In k-means the initial assignment of an instance (before the algorithm converges) is dependent on its nearest neighbour.
(B) In k-means clustering, k is learned and reflects the number of clusters.
(C) In k-medoids, the number of observed data points is equal to the number of clusters.
(D) In k-medoids the cluster representative (central point) is always an observed data point whereas in k-means this is not the case.

Question 15
Which description does not apply to the parameter k in the k-NN algorithm?

(A) It is the number of classes for the classification problem.
(B) It influences the smoothness of the decision boundary.
(C) It determines how many neighbours are considered for classifying a new example.
(D) It controls how well the model fits the training data.

Question 16
When using k-means on data that has circular properties, what is a possible undesirable outcome?

(A) The algorithm doesn't converge.
(B) The algorithm gets stuck in a local optimum.
(C) Different results on each different run of the algorithm.
(D) Cluster centers move to the circular data.

Question 17
Say you need to distribute 100 balls over 5 boxes. Explain in what situation the entropy of the distribution is the highest.

(A) When all boxes contain the same number of balls.
(B) When a single box contains all balls.
(C) When all boxes contain different numbers of balls.
(D) When the number of balls in each box is lg(100)

Question 18
What two criteria are being balanced in a typical SD quality measure for binary classification?

(A) The false positive rate and the false negative rate.
(B) The number of positives within the subgroup, and the size of the subgroup.
(C) The unusualness of the distribution of the target, and the size of the subgroup.
(D) The Weighted Relative Accuracy and the information gain.

Story B: Maximally Informative k-Itemsets

Consider the following dataset of binary attributes:

A	B	C	D
1	0	0	1
1	0	0	1
1	1	1	0
1	1	1	0
1	1	0	0
0	1	0	0
0	0	1	1
0	0	1	1

For answering the following questions, you may need to consult the following table of entropies.

p	$H(p)$
0	0
1/8	0.54
2/8	0.81
3/8	0.95
4/8	1
5/8	0.95
6/8	0.81
7/8	0.54
1	0

Question 19 (Story B)
What is the entropy of each of the four attributes over the entire dataset of Story B?

(A) $H(A) = 0.625, H(B) = 0.5, H(C) = 0.5, H(D) = 0.5$

(B) $H(A) = 0.95, H(B) = 1, H(C) = 1, H(D) = 1.$

(C) $H(A) = 0.95, H(B) = 0, H(C) = 0, H(D) = 0.$

(D) $H(A) = 0.287, H(B) = 1, H(C) = 1, H(D) = 1$

Question 20 (Story B)
Which itemset(s) is/are a miki with k = 2?

(A) $\{B, C\}$.

(B) $\{B, D\}$.

(C) $\{B, C\}$ and $\{C, D\}$..

(D) $\{B, C, D\}$

Question 21
Which small logical units do data warehouses keep large amounts of information in?

A. Data Storage
B. Data Marts
C. Access Layer
D. Data Miners

Question 22
What does the access layer help users to do?

A. Store Data
B. Analyse Data
C. Clean Data
D. Retrieve Data

Question 23
What type of data formats do conventional database systems use?

A. Highly Denormalized
B. Highly Normalized
C. A and B
D. None of the above

Question 24
Which of the following does the active data warehouse architecture include?

A. At least 1 data mart
B. Data that can be extracted from internal and external sources
C. Real-time updates
D. All of these

Question 25
A data warehouse is

A. Contains the world's data
B. Ready to be updated by the end-users
C. Organized around important subject areas
D. None of these

Question 26
A data mart is designed for the optimization of the performance for well-defined and predicable uses.

A. True
B. False

Question 27
Successful data warehousing needs a formal program in total quality management (TQM) to be implemented.

A. True
B. False

Question 28
When are dimensions conformed?

A. When they can be compared mathematically
B. When they are either the same or one is a subset of another.
C. When they have different values
D. When they are labeled differently

Question 29
What is reconciled data?

A. Data stored in one operational system
B. Data stored in the various operational systems available in the organization
C. Current data intended to be the single source for all decision support systems
D. Data that has been selected for end-user applications

Question 30
What is the system of Data Warehousing mostly used for?

A. Data Integration and Data Mining
B. Data Mining and Data Storage
C. Reporting and Data Analysis
D. Data Cleaning and Data Storage

Question 31
What is computing in data warehouses often referred to as?

A. OLAP
B. OLAT
C. OLTP
D. OLTA

Question 32
When does data staging occur in data warehousing?

A. A periodic process reads data from sources.
B. A periodic process stores data gotten from sources.
C. A periodic process mines data from sources
D. None of the above

Question 33
What is the combination of facts and dimensions sometimes called?

A. Physical Schema
B. Star Schema
C. Dimension Model
D. Denormalizing Modeling

Question 34
What does the typical Extract, Transform, Load (ETL)-based data warehouse use to house its key Functions?

A. Staging
B. Access Layers
C. Data Integration
D. All of the above

Question 35
A density-based clustering algorithm can generate non-globular clusters.

Question 36
Which of the following is a drawback in manual filling of missing values?

A. Incorrect data
B. Redundancy in data
C. Complex data
D. Too many missing values in the data

Question 37
Which of the following attribute values have a meaningful order but no information about the magnitude between successive values?

A. Nominal
B. Binary
C. Ordinal
D. Numeric

Question 38
TRUE/FALSE: A density-based clustering algorithm can generate non-globular clusters.

Question 39
TRUE/FALSE: Our use of association analysis will yield the same frequent itemsets and strong association rules whether a specific item occurs once or three times in an individual transaction

Question 40
TRUE/FALSE: The k-means clustering algorithm that we studied will automatically find the best value of k as part of its normal operation.

SECTION B [Answer all questions in this section: Total Marks = 80]

QUESTION TWO

2.1 KDD is really a good example of convergence of technologies where disciplines like statistics, graphics, mathematical and other analytical tools support KDD. Discuss with an example. (10)
2.2 Explain briefly any five data pre-processing approaches. (10)

QUESTION THREE

3.1 What is confusion matrix? Discuss various classification metrics along with their mathematical formulas (8)
3.2 The algorithm that we used to do association rule mining is the Apriori algorithm This algorithm is efficient because it relies on and exploits the Apriori property. What is the Apriori property? (2)
3.3 A database has 4 transactions, shown below.

TID	Date	items_bought
T100	10/15/04	{K, A, D, B}
T200	10/15/04	{D, A, C, E, B}
T300	10/19/04	{C, A, B, E}
T400	10/22/04	{B, A, D}

Assuming a minimum level of support min_sup = 60% and a minimum level of confidence min_conf = 80%:

(a) Find all frequent itemsets (not just the ones with the maximum width/length) using the Apriori algorithm. Show your work—just showing the final answer is not acceptable. For each iteration show the candidate and acceptable frequent itemsets. You should show your work similar to the way the example was done in the PowerPoint slides. (5)

(b) List all of the strong association rules, along with their support and confidence values, which match the following metarule, where X is a variable representing customers and $item_i$ denotes variables representing items (e.g., "A", "B", etc.).

$$\forall x \in transaction,\ buys(X, item1) \wedge buys(X, item2) \Rightarrow buys(X, item3)$$

Hint: don't worry about the fact that the statement above uses relations. The point of the metarule is to tell you to only worry about association rules of the form $X \wedge Y \Rightarrow Z$ (or $\{X, Y\} \Rightarrow Z$ if you prefer that notation). That is, you don't need to worry about rules of the form $X \Rightarrow Z$. (5)

QUESTION FOUR

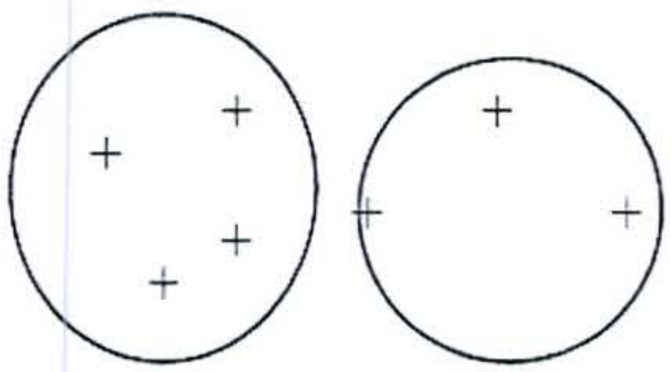
4.1. Outline the main differences between classification and clustering (6)

4.2. With the aid of examples describe at least three clustering methods (6)

4.3. Discuss the basic difference between the agglomerative and divisive hierarchical clustering algorithms and mention which type of hierarchical clustering algorithm is more commonly used

(4 Marks)

4.4. Are the two clusters shown below well separated? Answer: Yes or No and then in one or two sentences justify your answer. (4)



QUESTION FIVE

5.1. What are the differences between a data warehouse and a data mart? (2)

5.2. For a supermarket chain consider the following dimensions: product, store, time, promotion. The schema contains a central fact table, sales facts with three measures: unit_sales, dollars_sales, and dollar_cost. Design a star schema for this application.

Calculate the maximum number of fact table records for a warehouse with the following values:

Time period: 5 years

Store: 300 stores reporting daily sales three dimensions:

Product types: 40,000 products in each store (about 4000 sell in each store daily)

(6)

5.3. Illustrate how the supermarket can use clustering methods to improve sales (6)

5.4. Consider a data warehouse for a hospital, where there are three dimensions: (1) Doctor (2) (3) Time; and two measures: (1) Count and (2) Fees. (6)

END OF QUESTION PAPER