Mid-Term Exam

- Due Mar 4 at 5pm
- Points 50
- Questions 42
- Available Mar 4 at 2pm - Mar 4 at 6pm 4 hours
- Time Limit 75 Minutes

Instructions

- The exam on **modules 1, 2, 3, 4, 5, and 6.**
- The exam will be available on **Monday March 04, 2024 from 2:00 PM to 6:00 PM.**
- You need to answer **38 MCQs** with **1 point** for each + **4 Short questions** with **3 points** for each.
- You will have only **75 minutes** to complete your exam in **one sitting.**

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	75 minutes	35 out of 50 *

* Some questions not yet graded

ⓘ Correct answers will be available on Mar 5 at 6:30pm.

Score for this quiz: 35 out of 50 *
* Some questions not yet graded
Submitted Mar 4 at 4:10pm
This attempt took 75 minutes.

First Part: MCQs

Question 1
1 / 1 pts

Data mining activities can be subdivided into two major investigation streams , which are:

- Interpretation and Prediction

- Interpretation and Sampling
- Forecast and Prediction
- Sampling and Forecast

Question 2

1 / 1 pts

Which is the Application of Data Mining?

- Fraud Detection
- None
- Both
- Risk Analysis

Question 3

1 / 1 pts

Which of the following is a new trend in data mining?

- Invisible data mining
- All the three
- Web mining
- Scalable data mining methods

Question 4

1 / 1 pts

_____is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions.

- Data mart
- Business intelligence
- Business information warehouse
- Best practice

Question 5

1 / 1 pts

Data mining, the extraction of hidden ..... information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

- predictive
- preventive
- proactive
- provocative

Question 6

1 / 1 pts

Which of the following is not among functionalities (tasks) of data mining?

- Classification
- Clustering
- Association
- Visualization

Question 7

1 / 1 pts

What is Data Mining?

- The automated process of discovering patterns and relationships in an organization's data.
- The capability to drill down into an organization's data once a question has been raised.
- The process of performing trend analysis on the financial data of an organization.
- The setting up of queries to alert management when certain criteria are met.

Question 8

1 / 1 pts

Which of the following is not an objective of principal component analysis (PCA)?

- To reduce number of dimensions
-

To convert a set of observations of possibly uncorrelated variables into a set of values of linearly correlated variables

- To reduce attribute space from a larger number of variables to a smaller number of variables
- To identify new meaningful underlying variables

Question 9

1 / 1 pts

Estimated procedures can become rather complex and time-consuming for a large dataset with a high percentage of_____.

- Training data
- testing data
- resulting data
- missing data

Question 10

1 / 1 pts

Which of the following is a dimension reduction technique?

- All the three
- Stratified Sampling
- Box plot
- Principal component analysis

Question 11

1 / 1 pts

Data by itself is not useful unless

- It is properly stated
- It is processed to obtain information
- It is massive
- It is collected from diverse sources

Question 12

1 / 1 pts

The purpose of feature selection, also called _____.

- feature compression
- feature reduction
- feature normalization
- feature denormalization

Question 13

1 / 1 pts

The formula dist (x_i, x_k)= $\sqrt[q]{\Sigma|x_{ij} - x_{kj}|^q}$ shows:

- Cosine distance
- Euclidean distance
- Minkowski distance
- Manhattan distance

Question 14

1 / 1 pts

Continuous attributes are numerical attributes that assume an uncountable _____of values.

- first
- non-zero
- infinity
- zero

Question 15

1 / 1 pts

Which of the following is correct formula for accuracy of classifier?

- Accuracy = (TP + TN)/P
- Accuracy = N/(TP + TN)
- Accuracy = (TP + TN)/All

- Accuracy = (FP + FN)/All

Question 16

1 / 1 pts

The F-Measure is equal to zero if all the predictions are _____

- Incorrect
- Correct
- Partially incorrect
- Partially correct

IncorrectQuestion 17

0 / 1 pts

In data mining, what is the purpose of Interpretation?

- to identify irregular patterns in the data
- to express the rules and criteria for easy understanding
- All the three statements
- to determine useful patterns in the data

Question 18

1 / 1 pts

In weighted F-measure of precision and recall F(β), the value of β belongs to:

- [0, ∞)
- [0,1]
- [0, 1)
- [-1, 1]

Question 19

1 / 1 pts

Typically, classification matrix considers:

- Predicted Class

- Actual Class and Predicted Class
- Actual Class
- Target class

Question 20

1 / 1 pts

On which learning methods the Data Mining method is based?

- inductive learning methods
- deductive learning methods
- comprehensive learning methods
- basic learning methods

Question 21

1 / 1 pts

Among the following which method guarantees that each observation of the dataset appears the same number of times in the training set and exactly once in the test set.

- Cross Validation
- Holdout method and Repeated Random Sampling
- Repeated Random Sampling
- Holdout method

Question 22

1 / 1 pts

If the instances belongs to more than two classes then the classification is called as _____

- Binary Classification
- Multiclass Classification
- Double Classification
- High Classification

Question 23

1 / 1 pts

All of the following steps are part of Naïve Bayes method except:

- Express the probability as the product of p(x1|y) x p(x1|y) ... p(xn|y)
- Assign that class to the old record D.
- Determine what classes they all belong to and which is more prevalent
- Find all the other records where the predictor values are same

Incorrect Question 24

0 / 1 pts

Which of the following is not true for Bayes model for classification?

- All the records are used instead of relying on just the matching records
- Naïve Bayes classifiers are highly scalable
- Numerical variables need not to be converted into categorical
- Predictors should also be categorical

Question 25

1 / 1 pts

Naïve Bayes formula works well for-

- Classification
- Clustering
- Prediction
- Association

Question 26

1 / 1 pts

Rule-based Classification models are used to generate _____that allow the target class of future examples to be predicted.

- a set of misclassified variables
- a set of rules
- a set of targeted results

a set of predicted variables

Question 27

1 / 1 pts

Decision Trees or Association Rules are also called as?

knowledge discovery in databases

machine learning

data mining

All the three

Question 28

1 / 1 pts

Which of the following is a basis of Naïve Bayes method?

Regression

Pivot Table

Pie Chart

Conditional Probability

IncorrectQuestion 29

0 / 1 pts

In building a rule-based classifier, ------- use a function called One Rule function.

FOIL

C4.5rules algorithm

Indirect method

All the three

Question 30

1 / 1 pts

-------- are the strategies, in which each record is covered by at least one rule.

Mutually exclusive rules

- Not exhaustive rules
- Not mutually exclusive rules
- Exhaustive rules

Question 31

1 / 1 pts

In Logistic regression technique, input features can be

- Quantitative
- Qualitative
- Only numeric
- Quantitative and Qualitative

Question 32

1 / 1 pts

K- Nearest Neighbor Classifier is know as:

- All the three
- Local classifier
- Instance-based learner
- Lazy learner

Question 33

1 / 1 pts

One of the important characteristics of K-Nearest Neighbor Classifier is:

- These classifiers can handle the missing values
- They usually work well in the presence of irrelevant and redundant attributes
- They usually make their predictions based on local information
- All the three

Question 34

1 / 1 pts

How to determine the class label of a test example when using the K-Nearest Neighbor?

- All the three
- Take the majority vote of class labels among the all k nearest neighbors
- Weight the vote according to distance to reduce the impact of K neighbors

Choose a right method for using class labels of K nearest neighbors to determine the class label of unknown record

Question 35

1 / 1 pts

Why data preprocessing is high recommended when using K-Nearest Neighbor Classifier?

- Proximity computations normally require the presence of all attributes
- To avoid any situation, in which one of the attributes can dominate our distance measure
- All the three
- To let the classifier handling missing values in both the training and test sets

Question 36

1 / 1 pts

Logistic regression is a _____ regression technique that is used to model data having a _____outcome

- Nonlinear, numeric
- Nonlinear, binary
- Linear, binary
- Linear, numeric

Question 37

1 / 1 pts

Which of the following methods do we use to best fit the data in Logistic Regression?

- Least Square Error
- Euclidean distance
- Maximum Likelihood

- Jaccard distance

Question 38

1 / 1 pts

Function which is used to bound the probability of x between 0 and 1?

- Cosine
- Sine
- Sigmoid function
- Log function

Second Part: Short Questions

Question 39

Not yet graded / 3 pts

Describe ***one data mining's issue*** that, in your view, may have a strong impact on the market and on society. Briefly, discuss ***how to approach such an issue***.

Your Answer:

One of the data mining issue that have the strong impact on the market and the society is misuse of the personal information. This leads to the security problem and privacy breaches. This problem is occured when the data is obtained for mining for marketing purposes is being misused which harm the society. To approach this issue it is important to prioritize the data privacy and security measures, like the organizations shold establish the data privacy policies, limiting the access to personal data. This can also be controlled by providing education and awareness programs within the organizations.

Question 40

Not yet graded / 3 pts

What do we mean by ***pruning*** the decision tree? Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. What advantage does (a) have over (b)?

Your Answer:

When we build the decision tree, we can observe number of trees that have noise in training of the data.

If the the purning is done afterwards, so that the training set can be classified perfectly and also instead of purning the entire tree the single node is purned first. This helps to simplify the process of the decision making which helps to enhance the performance of the model. First converting to rules and then purning also helps to generate refined and interpretable rules.

Question 41
Not yet graded / 3 pts

Consider a training set that contains **32 positive** examples and **224 negative** examples. For the of the following candidate rule,

***R*1: *A* → + (covers 8 positive and 24 negative examples),**

Determine its FOIL's information gain.

(**Hints**: 1) You can type the logarithm of base 2 as lg.

2) lg (x/y) = lg x - lg y and lg (xy) = lg x + lg y)

Your Answer:

p_0 = number of positive data instances

n_0 = number of negative data instances

$p_0 = 32$, $n_0 = 224$.

$p_1 = 8$, $n_1 = 24$, $p_0 = 32$, $n_0 = 224$

information gain for R1 = 8 [log (8 / 32) - log (32 / 256)] = 8

Question 42
Not yet graded / 3 pts

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

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

Indicate your reasoning if you think there may be some ambiguity in some cases.

Example: Age in years.

Answer: Discrete, quantitative, ratio.

A. Number of students enrolled in a class.
B. Daily user traffic volume at YouTube.com (i.e., number of daily visitors who visited the Web site).

Your Answer:

A. Discrete, quantitative, ratio.

the no.of students enrolled is a whole number which makes it discrete and also it represents the numerical quantity which states it is quantitative and it is ratio.

B. Discrete or Continuous, quantitative, ratio.

The daily traffic is either discrete or continuous, if the traffic is measured as whole number then it is discrete or if it is continuous then it is continuous.

Quiz Score: 35 out of 50
* Some questions not yet graded