Zagazig University Faculty of Computers and Informatics Year. fourth Year Subject Name. Data mining and machine learning (IS405) Department. Information systems department	Final Exam	Date. 11/6 / 2024 Time Allowed. 3 Hour. No. of Pages. 5 No. of Questions. 2 Model. 1 Total. 60 Marks Second term exam

Answer All Questions

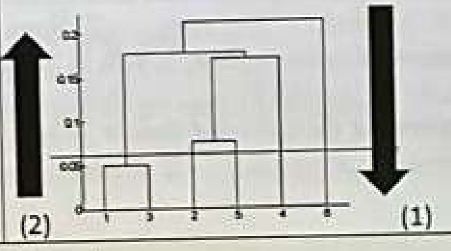
Question [1]: multiple choice questions (60 Marks)

1. **How many clusters in the following dendrogram?**
   A. 3 B. 6 C. 4 D. 5
2. **According to the following dendrogram the distance between the clusters 1 and 3 is**
   A. 0.05 B. 0.1 C. 0.15 D. 0.2
3. **According to the following dendrogram the bottom up approach (2) refer to ....**
   A. Partitional Clustering B. K-Means Clustering
   C. Agglomerative Clustering D. Divisive Clustering
4. **According to the following dendrogram the up down approach (1) refer to ....**
   A. Partitional Clustering B. K-Means Clustering
   C. Agglomerative Clustering D. Divisive Clustering
5. **Start with the points as individual clusters**
   A. Partitional Clustering B. K-Means Clustering
   C. Agglomerative Clustering D. Divisive Clustering
6. **According to the following distance matrix the first cluster include:**
   A. P3, p6, p1 B. p3, p6 C.p2, p5 D.p2, p4
7. **According to the following distance matrix the second cluster include:**
   A. P3, p6, p1 B. p3, p6 C.p2, p5 D.p2, p4
8. **To update the distance matrix using MIN[dist(p2, p5), p1]is..........**
   A. 0.34 B. 0.15 C. 0.23 D. 0.14
9. **To update the distance matrix using MAX[dist(p2, p5), p4] is ......**
   A. 0.29 B. 0.23 C. 0.20 D. 0.14
10. **To update the distance matrix using Average [dist(p2, p5), p4] is ......**
    A. 0.293 B. 0.235 C 0.285 D 0.245

Distance matrix

	P1	P2	P3, p6	P4	P5
P1	0				
P2	0.23	0			
P3, p6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0

Dendrogram



11. The more popular hierarchical clustering technique is:
    A. Agglomerative B. Divisive C. K-means D. Partitional Clustering
12. **Two documents are represented by the following two vectors:**
    d1 = (3 2 0 5 0 0 0 2 0 0)
    d2 = (1 0 0 0 0 0 0 1 0 2) Using cosine similarity, cos (d1, d2) is
    A 0.658 B 0.315 C 0.875 D 0.946
13. In problem **(12)** cosine dissimilarity =
    A. 0.064 B. 0.178 C. 0.685 D. 0.045

14. **Sequential ensemble techniques generate base learners in a sequence, e.g.,**

A. KDD B. Random Forest C. KNN D. Adaptive Boosting

15. In **parallel ensemble techniques**, base learners are generated in a parallel format, e.g.,

A. Ada boost B. Gradient Boosting C. XG boost D. Random Forest

16. **..............Construct a set of base classifiers learned from the training data then Predict class label of test records by combining the predictions made by multiple classifiers (e.g., by taking majority vote)**

A. KNN B. Ensemble Methods C.KDD D. Decision tree

17. **.....................Can be used to solve the problem of overfitting in deep decision tree**

A. KNN B. Hierarchical clustering C. Random Forest D. Agglomerative

18. **Synonym for data mining is Select one:**

A. Data Warehouse B. KDD
C. Business intelligence D. OLAP

19. **Simple Matching Coefficients and Jaccard Coefficients to**

x= 1000000000 y= 0000001001 =............And ....

A . 0.7 and 0 B. 0 and 0.7 C. 0.7 and 2 D. 0.6 and 4

20. **Distance between 2 objects X and Y using Euclidean Distance where X= (3, 1) and Y= (5,1) is....................**

A. 3.62 B. 7.324 C. 2 D.4.326

21. **Hamming distance to p1 = 10101011 and p2= 10011 10 1 is ....**

A. 3 B. 4 C. 5 D. 2

22. **...............is a metric for comparing two binary data strings and is used for error detection or error correction when data is transmitted over computer networks.**

A. Euclidean Distance B. Manhattan distance C. Minkowski Distance D. Hamming distance

23. **Numerical measure of how alike two data objects are**

A. Dissimilarity measure B. Proximity measure
C. Similarity measure D. Distance

24. **In k-means algorithm the number of centroids for two clusters=........**

A. 3 **B. 2** **C.1** **D.4**

25. **K-means has problems when the data contains**

A. Outliers B. Noise C. Error D. Inconsistence

26. **K-means has problems when clusters are of different**

A. Sizes B. Densities C. Non-globular shapes D. All the previous

27. **.............attributes are a special case of discrete attributes**

A. Binary B. Numeric C. Unary D. None of the previous

28. **What is the median of the sample 8, 7, 6, 9, 5, 3, 4 ?**

A .5 B. 6 C. 8 D. 9

29. **Which of the following data mining task is known as Market Basket Analysis? Select one:**

A. Outlier Analysis B. Regression C. Classification D. Association Analysis

30. **Which of the following are descriptive data mining activities? Select one:**

A. Recommendation B. Classification C. Clustering D. Regression

31. **Other names for attribute except for:**

A. Variable B. Characteristic C. Field D. Design

32. **An attribute is ....**

A. Dataset B. A property or characteristic of an object
C. A collection of data objects D. An observation

33. **Which of the following is/are Prediction methods**

A. Classification B. Regression C. clustering D. both A and B

4. The ________ refers to extracting knowledge from larger amount of data.

A. data abstraction. B. data mining. C. database. D. data warehouse

35. ....routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

A. Data cleaning B. Data Integration C. Data Reduction D. Data Transformation

36. Nominal and ordinal attributes can be collectively referred to as________ attributes Select one:

A. Perfect B. Quantitative C. Qualitative. D. Optimized

37. Predicting tumor cells as benign or malignant is an example of.........

A. Clustering B. Regression C. Classification D. Anomaly detection E. Association

38. ....A set of nested clusters

A. Partitional Clustering B. Hierarchical Clustering C. K-means D. KNN

39. Error or outlier data is known as ________ Select one:

A. Missing data B. Inconsistence C. Changing data D. Noisy data

40. Combining two or more attributes (or objects) into a single attribute (or object)

A. Sampling B. Aggregation C. Discretization D. Binarization

41. ....is the process of converting a continuous attribute into an categorical attribute

A. Sampling B. Aggregation C. Discretization D. Binarization

42. ....is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

A. Attribute transform B. Dimensionality Reduction

C. Principal Components Analysis D. None of all

43. ....data provides the information that identifies the location of features and boundaries on Earth.

A. Spatial B. Temporal C. Graphs D. Time series

44. .... from data quality problems except :

A. Noise and outliers B. Missing values C. Duplicate data D. Relevance

45. ....Split the data into several partitions; then draw random samples from each partition

A. Random Sampling B. Sampling without replacement

C. Sampling with replacement D. Stratified sampling

46. Two fundamental goals of Data Mining are ____ ____.

A. Analysis and Description B. Data cleaning and organizing the data

C. Prediction and Description D. Analysis and Prediction

47. Data mining is -------------

A. An extraction of explicit, known and potentially useful knowledge from information.

B. A non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data.

C. An essential process where intelligent methods are applied to extract data patterns that is also referred to database.

D. Is an essential process where intelligent methods are applied to extract data that is also referred to data sets.

48. Which of these a method to perform data transformation?

A. Data compression B. Normalization C. Filling in missing data D. Dimensionality reduction

49. Which type of data is a sequential data recorded in specific time intervals?

A. Time Series Data B. Spatial Data C. Sequence Data D. Transactional Data

50. Variables whose measurement is done in terms such as weight, height and length are classified as

A. flowchart variables B. measuring variable C. continuous variables D. discrete variables

51. The observation which occurs most frequently in a sample is the

A. median B. mean deviation C. standard deviation D. mode

52. Which of the following is not a data pre-processing methods Select one:

A. Data Visualization B. Data Discretization C. Data Cleaning D. Data Reduction

53. Dimensionality reduction reduces the data set size by removing ………………… Select one:

A. Composite attributes B. Derived attributes C. Relevant attributes D. Irrelevant attributes

54. Nominal and ordinal attributes can be collectively referred to as_________ attributes Select one:

A. Perfect B. Quantitative C. Qualitative. D. Optimized

55. Predicting tumor cells as benign or malignant is an example of………

A. Clustering B. Regression C. Classification D. Anomaly detection E. Association

56. The difference between supervised learning and unsupervised learning is given by Select one:

A. unlike unsupervised learning, supervised learning needs labeled data

B. unlike unsupervised learning, supervised learning can be used to detect outliers

C. there is no difference

D. unlike supervised leaning, unsupervised learning can form new classes

57. Identify the example of Nominal attribute Select one:

A. Temperature B. Eye color C. Mass D. Salary

58. Normalization include:

A. min-max B. z-score C. Both (A and B) D. Discretization

59. What is an example of data quality problems?

A. Noise B. Outliers C. Duplicate Data D. All of the previous

60. ________analysis divides data into groups that are meaningful, useful, or both.

A. Cluster. B. Association. C. Classification. D. Relation

Question [2]: True or false (30 marks)

61. In Divisive Clustering, at each step merge the closest pair of clusters until only one cluster left.

62. In decision tree, Entropy is the only measures of node Impurity.

63. Manhattan is a generalization of Euclidean Distance.

64. If we have 2 objects X and Y and they have nominal attributes we use the property if X=Y then the distance between X and Y = 0

65. From Common Properties of a Distance if d (x, y) not equal d (y, x) for all x and y the they are Symmetry.

66. Principal Components Analysis (PCA) is used for Dimensionality Reduction.

67. Data post processing is perhaps the most laborious and time-consuming step in the overall knowledge discovery process.

68. The purposes of Dimensionality Reduction are to reduce the amount of time and memory required by data mining algorithms and allow the data to be more easily visualized.

69. Military ranks is an example of ordinal attributes?

70. Jaccard distance:

$$JDist(X,Y) = 1 - JSim(X,Y)$$

71. Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model .

72. Homogenous base learners refer to base learners of the same type, with similar qualities.

3. In Ensemble methods, Boosting works by training a large number of strong learners arranged in a parallel pattern and then combining them to optimize their predictions.

74. In bagging technique, Bootstrapping is a sampling technique where samples are derived from the whole population using the replacement procedure.

75. Random forest is Heterogeneous base learners.

76. In Random forest, the majority voting in regression we use mean.

77. In Random Forest the forest it builds, is an ensemble of decision trees.

78. Using a sample will work almost as well as using the entire data set, if the sample is representative.

79. Regression is a descriptive data mining task .

80. In statistics, standardization refers to subtracting off the means and dividing by the standard deviation.

81. In Dissimilarity measure Upper limit =1.

82. Dissimilarity measure is lower when objects are more alike.

83. Binarization maps a binary variable into one or more continuous or categorical attributes.

84. Sampling is the main technique employed for data reduction.

85. Traditional Techniques may be unsuitable for extract information because of Enormity of data.

86. Inter-cluster distances are maximized.

87. In K-means it is stopped when there is a change in the cluster.

88. Discrete attributes are often represented using real numbers.

89. Dataport and UCI are From sources of data sets.

90. To detect fraudulent usage of credit cards, the Outlier analysis data mining task should be used.

Dr. Gawaher Soliman

With my best wishes