D:\exam01\logo_red.pngVISHNU INSTITUTE OF TECHNOLOGY (AUTONOMOUS)

Mid – II Examinations

**Data warehousing and Mining (AIML and AIDS)**

**II-II**

Unit 3

1. Describe the data classification process with a neat diagram. How does the Naive Bayesian classification work? Explain L1,C3, 6 Marks
2. What is the difference between overfitting and under fitting? Explain their relation with training error and generalization error. L2, C3, 6 Marks
3. Why is Naïve bayes classifier called Naïve? What are the characteristics of naïve bayes classifier? L2,C3, 6Marks
4. What do you mean by confusion matrix? How are the performance metrics like accuracy, error rate, precision and recall calculated? Explain with an example L1, C3 6 Marks
5. Write a short note on below model evaluation techniques: L1, C3, 6 Marks
   1. Holdout
   2. Random Sampling
   3. Two-fold cross validation
   4. K-fold cross validation
   5. Bootstrap
   6. Leave one out approach
6. What do you mean by Training errors and Generalization errors? If a model has very less training error but very large generalization error, the model is known as? L3, C3, 6 Marks
7. Consider the below table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play Golf** |
| 0 | Rainy | Hot | High | FALSE | No |
| 1 | Rainy | Hot | High | TRUE | No |
| 2 | Overcast | Hot | High | FALSE | Yes |
| 3 | Sunny | Mild | High | FALSE | Yes |
| 4 | Sunny | Cool | Normal | FALSE | Yes |
| 5 | Sunny | Cool | Normal | TRUE | No |
| 6 | Overcast | Cool | Normal | TRUE | Yes |
| 7 | Rainy | Mild | High | FALSE | No |
| 8 | Rainy | Cool | Normal | FALSE | Yes |
| 9 | Sunny | Mild | Normal | FALSE | Yes |
| 10 | Rainy | Mild | Normal | TRUE | Yes |
| 11 | Overcast | Mild | High | TRUE | Yes |
| 12 | Overcast | Hot | Normal | FALSE | Yes |
| 13 | Sunny | Mild | High | TRUE | No |

Ramu wants to play golf. And let today’s weather is (Outlook=Sunny, Temperature=Hot, Humidity=Normal, Windy=False). Is it favorable to play golf today?

Unit 4

1. Apply Apriori algorithm for frequent itemset mining in transactional databases. Apply these algorithms to the following data:

TID LIST OF ITEMS

1 Bread, Milk, Sugar, TeaPowder, Cheese, Tomato

2 Onion, Tomato, Chillies, Sugar, Milk

3 Milk, Cake, Biscuits, Cheese, Onion

4 Chillies, Potato, Milk, Cake, Sugar, Bread

5 Bread, Jam, Mik, Butter, Chilles

6 Butter, Cheese, Paneer, Curd, Milk, Biscuits

7 Onion, Paneer, Chilies, Garlic, Milk

8 Bread, Jam, Cake, Biscuits, Tomato L3, C4, 12 Marks

1. A. What is Association rule mining? Briefly describe the criteria for classifying association rules. L1, C4, 6 Marks
2. What are the drawbacks of Apriori Algorithm? How they are overcome using FP Growth algorithm. L2, C4, 6 Marks
3. Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, explain it with the following table: L3, C4, 12 Marks

TID List of items

001 Milk, dal, sugar, bread

002 Dal, sugar, wheat,jam

003 Milk, bread, curd, paneer

004 Wheat, paneer, dal, sugar

005 Milk, paneer, bread

006 Wheat, dal, paneer, bread

1. A database has four transactions. Let min\_sup=60% and min\_conf=80%

TID date items\_bought

100 10/15/2018 {K, A, B, D}

200 10/15/2018 {D, A, C, E, B}

300 10/19/2018 {C, A, B, E}

400 10/22/2018 {B, A, D}

i) Find all frequent items using Apriori.

ii) List all the strong association rules (with support ‘s’ and confidence ‘c’) matching the following meta-rule where X is a variable representing customers, and item i denotes variables representing items (e.g., “A”, “B”, etc.): Vx Є transactions, buys (X, item1) ^ buys (X, item2) =>buys (X, item3) [s,c]. L3, C4, 12 Marks

1. A. Discuss the applications of association analysis. L2, C4, 6 Marks

B. Explain the following terms:

a. Frequent itemset

b. Support

c. Confidence L1, C4, 6 Marks

1. A. Explain the following terms:
   1. Apriori principle
   2. Anti monotone property
   3. Support based pruning L1, C4, 6 Marks
2. A. Explain Fk-1 \*Fk-1 method for generating itemsets? Explain with and example.

L1, C4, 6 Marks

B. How is computational complexity of Apriori algorithm determined? Explain in brief.

1. How to find frequent item sets using candidate generation? Explain with algorithm

L1, C4, 12 Marks

1. How to find frequent item sets without candidate generation? Explain with algorithm
2. A. What are the advantages and disadvantages of Apriori and FP growth algorithms?

L1, C4, 6 Marks

B. Explain the terms

i) Maximal frequent itemset

ii)Closed itemset

iii)Closed frequent itemset L1, C4, 6 Marks

Unit 5

1. Consider five points {X1 , X2 , X3 , X4 , X5 } with the following coordinates as a two dimensional sample for clustering : X1 = ( 0.5,2.5 ); X2 = ( 0,0 ); X3 = ( 1.5,1 ); X4 = ( 5,1 ); X5 = (6,2 ) Illustrate the K-means partitioning algorithms using the above data set.

L3, C5, 12 Marks

1. A. What is Density based clustering? Describe DBSCAN clustering algorithm. L1, C5, 6 Marks

B. What is the difference between a cluster and an outlier? Write the practical applications of outlier analysis L2, C5, 6 Marks

1. A. Distinguish Agglomerative hierarchical clustering from divisive hierarchical clustering. What are the different methods for performing Agglomerative hierarchical clustering?

L2, C5, 6 Marks

B. what are the Key issues in hierarchical clustering? L1, C5, 6 Marks

1. Illustrate k-means clustering algorithm. Explain with an example to demonstrate the same

L2, C5, 12 Marks

1. A. Explain bisecting K-Means? How is it different from Simple K-Means. L2, C5, 6 Marks

B. What are the additional Issues of K-Means? L1, C5, 6 Marks

1. A. Write a brief note on K-means and Different Types of Clusters L1, C5, 6 Marks

B. Write the Strengths, Weaknesses, Time, and space complexity of K-Means?

L1, C5, 6 Marks

1. What are the methods for defining the proximity between the clusters? Explain Single- Link hierarchical clustering with example? L3, C5, 12 Marks
2. A. Write the strengths, weakness, time, and space complexity of hierarchical clustering?

L1, C5, 6 Marks

B. Explain center-based approach for density-based clustering? L1, C5, 6 Marks

1. A. what are different types of clustering techniques? L1 C5, 6 Marks

B. What are different types of clusters? L1, C5, 6 Marks

1. What is the difference between Single link, Complete Link, Group average and ward’s approaches for Agglomerative hierarchical clustering? L2, C5 12 Marks