EID902: INTRODUCTION to DATA MINING

QUESTION BANK

# Module 1: Introduction to Data Mining

1. What is Data Mining? Explain the process of Data Mining with a neat diagram.
2. Why to mine the data?
3. With a neat diagram explain the origins of data mining.
4. What are the different tasks of data mining? List and explain.
5. List and explain various motivating factors for data mining.
6. Define the following terms:
   1. Data objects
   2. Attribute
7. Discuss the different types of attributes and their properties.
8. List and explain different characteristics of a data.
9. Write a note on the following:
   1. Record data sets
   2. Graph data sets
   3. Ordered data sets
10. Write a note on Data Quality
11. Write a brief notes on the following:
    1. Similarity measures
    2. Dissimilarity measures
12. Enlist various data pre-processing techniques
13. What is aggregation? What is its purpose? Explain.
14. What is sampling? What is its purpose? Explain the different sampling techniques.
15. What is discretization? What is its purpose? Explain the different discretization techniques.
16. What is binarization? What is its purpose? Explain the different binarization techniques.
17. What is curse of dimensionality? What is the purpose of dimensionality reduction? Explain the different dimensionality reduction techniques.
18. What is feature subset selection? Explain.
19. What is feature creation? What is its purpose? Explain the different feature creation techniques.
20. Distinguish between noise and outliers.
21. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors. x = 0101010001, y = 0100011000
22. For the following vectors, x and y, calculate the indicated similarity or distance measures.
    1. x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlation, Euclidean.
    2. x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard
    3. x = (0, −1, 0, 1), y = (1, 0, −1, 0) cosine, correlation, Euclidean
    4. x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard
    5. x = (2, −1, 0, 2, 0, −3), y = (−1, 1, −1, 0, 0, −1) cosine, correlation

# Module 2: Classification

1. Define classification and with a neat diagram explain the general approach for building classification model.
2. Enlist different classification techniques.
3. Explain the general structure of Hunt’s algorithm. Discuss the design issues of decision tree induction.
4. Explain the methods for expressing test conditions.
5. Discuss the various impurity measures.
6. Explain the process of finding the Best Split
7. List the merits and limitations of Decision Tree Based Classification
8. Define the following:
   1. Training errors,
   2. Test errors,
   3. Generalization errors
   4. Under fitting,
   5. Overfitting.
9. What are the reasons for model overfitting? What are the effect of multiple comparison procedure?
10. What are the different ways for estimating generalization errors? Explain.
11. Write a note on tree pre-pruning and post-pruning.
12. Explain various methods for handling the model evaluation.
13. What is Artificial Neural Networks (ANN)? Explain the basic architecture of perceptron.
14. Explain the learning algorithm for basic perceptron.
15. What is multi-layer neural network? Explain the learning in multi-layer neural network.
16. Write a note on Gradient Descent.
17. Describe various design issues of ANN. Describe various Characteristics of ANN.
18. Discuss the trends in Deep Learning.
19. What is k-NN? Explain the k-NN algorithm. What are the design issues in k-NN? How do you improve the k-NN efficiency?
20. Describe Bayes’ Theorem. Explain Naïve Bayes classifier.
21. Describe Bayesian Belief Networks for classification.
22. What is a rule based classifier? Discuss the properties of a good rule set. Describe various rule ordering schemes.
23. Write a note on direct classification rule building technique.
24. Describe the various rule evaluation measures.
25. Write a note on RIPPER.
26. List the advantages and limitation of rule based classifier.
27. What is support vector machine? How Learning occurs in SVM for solving the linear separable classes?
28. How learning occurs in SVM for solving the non-linearly separable classes? Explain.
29. Discuss the various characteristics of SVM.

# Module 4: Clustering

1. What is cluster analysis? What are the applications of cluster analysis?
2. What is clustering? What are the various types of clustering?
3. What is a cluster? What are the various types of clusters?
4. How the input data affects the clustering process? Explain.
5. What is k-Means algorithm? What are the characteristics of k-Means algorithm?
6. Discuss the design issues of k-Means algorithm.
7. Write a note on Bisecting k-Means algorithm.
8. Describe the hierarchical clustering. What are the main types of hierarchical clustering?
9. Describe the Agglomerative Clustering Algorithm.
10. Describe various measure for calculating the inter-cluster distance? Describe their strengths and weaknesses.
11. What are the various characteristics of hierarchical clustering methods?
12. What is DBSCAN? Describe the basic DBSCAN algorithm. Discuss the design issues of DBSCAN algorithm.
13. Write a note on unsupervised clustering measures?
14. How do you measure the cluster validity via correlation?
15. Write a note on supervised clustering measures?
16. Explain the BIRCH algorithm in brief.
17. Explain the CURE algorithm in brief.

# Module 5: Outlier or Anomaly Analysis

1. What are anomalies/outliers? What are the causes of anomalies?
2. What are the distinction between?
3. Noise and anomalies
4. Model based and Model free approaches
5. List the various anomaly detection techniques.
6. Write a note on
   1. Statistical approaches.
   2. Distance based approaches
   3. Density based approaches
   4. Clustering based approaches
   5. Reconstruction based approaches
   6. Information theoretic approaches
7. How does One-Class SVM Work?
8. Write a note on evaluation of anomaly detection approaches.

# Module 3: Association Analysis

1. Describe the Apriori algorithm for deriving the association rules.
2. Describe the F-P tree algorithm for deriving the association rules.
3. Solve the problems using the above algorithms.