**Question bank for Module 1 and Module 2**

1. What is Data Mining?
2. Identify the origins of Data Mining.
3. List the motivations for data mining.
4. What are the tasks of data mining?
5. Illustrate the applications of classification.
6. Illustrate the applications of clustering?
7. Demonstrate the applications of outlier analysis?
8. Clarify the applications of association rule mining?
9. Define the following:
   1. Object.
   2. Attribute.
10. Describe the classification of attributes.
11. Define
    1. Discrete valued attribute.
    2. Continuous valued attribute.
12. Enlist the important characteristics of data.
13. What is a market basket data? Explain.
14. Describe the sparse matrix?
15. Clarify the graph based data.
16. Describe the Spatial Data.
17. Define the term ‘temporal coherence’.
18. List the data quality issues.
19. How do you handle the missing data?
20. Define a metric.
21. Describe the Pearson correlation coefficient.
22. What is aggregation? What are the goals of aggregation in data mining?
23. How do you sample the data?
24. What is binarization?
25. What is transformation? List the different transformation techniques.
26. Elaborate the curse of dimensionality.
27. What are the decision tree induction design issues?
28. Define: Gini Index.
29. Give the definition for the Gain.
30. What is classification? List the different classifiers.
31. Explain the methods for expressing ordinal attribute test conditions.
32. Explain the methods for expressing nominal attribute test conditions.
33. Explain the methods for expressing continuous attribute test conditions.
34. What do you mean by descriptive modelling?
35. What do you mean by predictive modelling?
36. For the following data calculate the SMC: x={1,0,0,0,0,0,0,0,0,1}, y={0,0,0,0,0,0,1,0,1,0}
37. For the following data calculate the J: x={1,0,0,0,0,0,0,0,0,1}, y={0,0,0,0,0,0,1,0,1,0}
38. What is a document matrix? Explain.
39. Describe the general approach to classification task.
40. Explain the different methods for feature subset selection.

**Course Name: Data Mining.**

**Classification and Clustering--Question Bank**

**Note: Answer all the questions.**



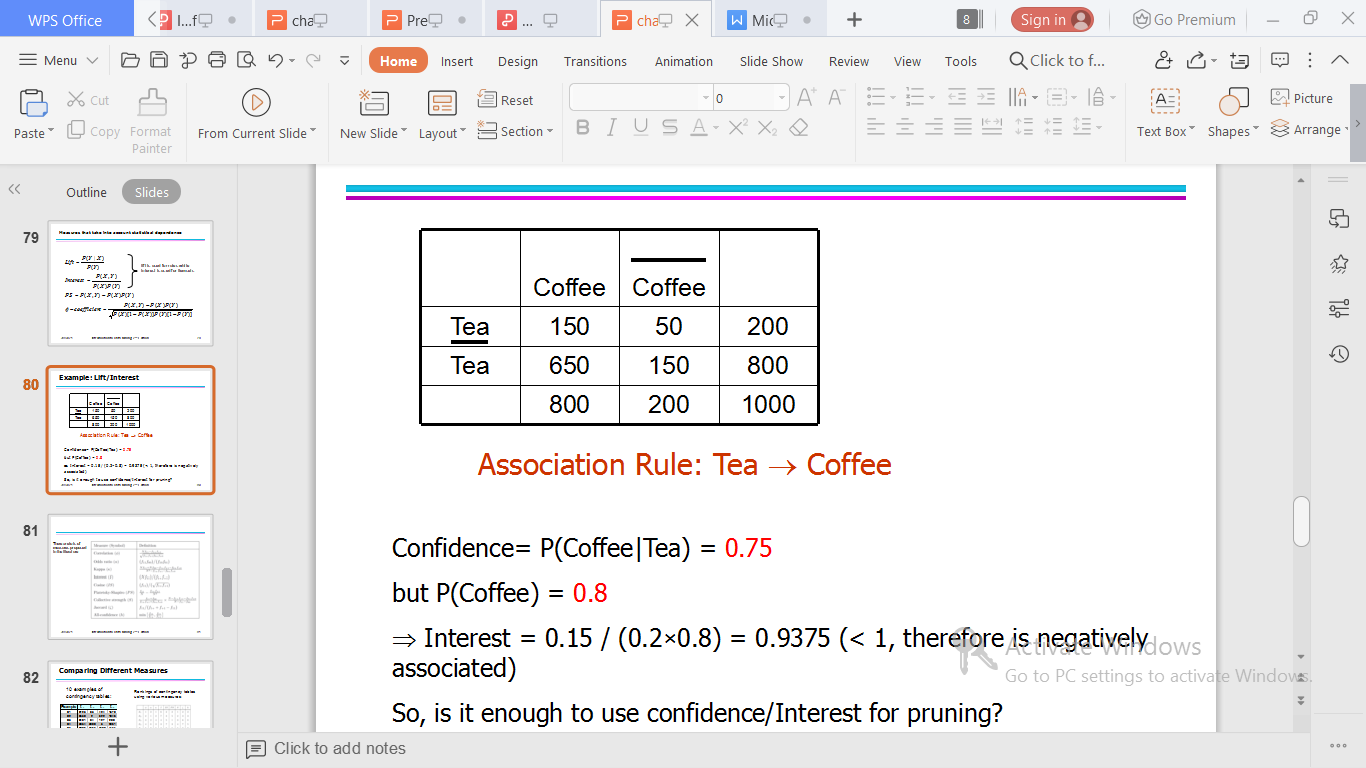
|  |  |  |
| --- | --- | --- |
| **Q#** | | **Question** |
|  |  | Describe the decision tree induction (**DTI**) algorithm. |
|  |  | What are the potential causes of model overfitting? |
|  |  | How do you determine the right complexity of the model? |
|  |  | Discuss the various strategies to handle overfitting in DTI. |
|  |  | Explain the k-fold cross-validation technique. |
|  |  | Explain the bootstrapping technique. |
|  |  | What are the characteristics of a rule set? Describe. |
|  |  | Discuss the sequential covering algorithm. |
|  |  | Use Naïve Bayesian classifier for labelling the example:  **Table 1: Loan defaulter dataset.**  **Xt= {Refund = Yes, Divorced, 120K}.** |
|  |  | Explain the learning algorithm for **Perceptron**. |
|  |  | Describe the learning process in the support vector machines **(SVM)** approach. |
|  |  | What is clustering? Discuss the different types of clustering. |
|  |  | For the **k-means** approach, enlist the common choices for proximity, centroid, and objective functions. |
|  |  | Describe the basic agglomerative hierarchical clustering (**AHC**) algorithm. |
|  |  | Discuss the **DBSCAN** algorithm. |
|  |  | Write a note on the **CURE** algorithm. |
|  |  | Write a note on the **BIRCH** algorithm. |
|  |  | A training set contains 100 positive and 400 negative examples. For each of the following candidate rules, determine the best rule using **FOILs** information gain measure.  **Table 2: Rule Set**   |  |  |  | | --- | --- | --- | | Rule | Positive examples covered | Negative examples covered | | r1 | 4 | 1 | | r2 | 30 | 10 | | r3 | 100 | 90 | |
|  |  | List the characteristics of the **k-NN** classifier. |
|  |  | Describe the topology generating algorithm of Bayesian belief networks **(BBN).** |

**Course Name: Data Mining.**

**Outlier analysis and** **Association Analysis**

**Question Bank**

**Note: Answer all the questions.**

1. What are anomalies/outliers? Describe.
2. Explain the different causes of anomalies.
3. Differentiate between the outliers and noisy points.
4. Explain the statistical anomaly detection techniques.
5. What are the strengths and weaknesses of statistical anomaly detection techniques?
6. Describe the density based anomaly detection techniques.
7. What are the strengths and weaknesses of distance based anomaly detection techniques.
8. How to detect the outliers using the clustering methods? Explain.
9. What is an auto-encoder? List its strengths and weaknesses.
10. How do you evaluate the performance of anomaly detection techniques?
11. Describe the following terminologies
    1. Support
    2. Confidence
12. Briefly describe the association rule mining task
13. Describe various frequent itemset generation strategies.
14. Explain apriori algorithm for frequent itemset generation.
15. List the different factors that affect the apriori algorithm complexity.
16. Table 1: People preference
17. For the following data shown in table 1, find out
    1. Lift
    2. Interest
18. Define a “sequence”.
19. Describe the sequential pattern mining.
20. Explain the Piatesky-Shapiro (PS) measure
21. Define FP Tree.