**True/false**:

|  |  |
| --- | --- |
| **T** | 1. In classification using decision trees, if the tree is fully completed in depth, it is more likely that it fits noise in the data hence become overfitting. |
| **T** | 1. K-mean assigns each instance, or case, to a unique cluster based on the distance that it has to the cluster center. |
| **T** | 1. The value of TFIDF is specific to a single document, whereas IDF value depends on the entire corpus. |
| **T** | 1. In text mining, if an association between two concepts has 7% support, it means that 7% of the documents had both concepts represented in the same document. |
| **T** | 1. The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. |
| **F** | 1. The equi-depth approach divides the range of attribute into a user-specified number of intervals each having the same width. |
| **F** | 1. An equi-width approach tries to put the same number of objects into each interval. |
| **T** | 1. In sentiment analysis, it is hard to classify some subjects such as news as good or bad, but easier to classify others, e.g., movie reviews, in the same way. |
| **T** | 1. The confusion matrix is a technique used to evaluate the results of classification, in terms of correctness and errors. |
| **F** | 1. Confidence does not change by itemset order change, whilst support does change when the itemset order change. |
| **F** | 1. K-mean defines the prototype in terms of the most representative point for a group of points. |
| **T** | 1. Root node has no incoming edges and zero or more outgoing edges. |
| **F** | 1. Internal node has one or more incoming and two or more outgoing edges. |
| **F** | 1. Leaf node has exactly one incoming and one outgoing edge. Each leaf node is assigned a class label. |
| **F** | 1. Tokenizing is the process of reducing multiple words to their base or root and then transforming the term-by-document matrix to a manageable size. |

**MCQ:**

Use these tables to answer next two questions:

|  |  |
| --- | --- |
| **Single Item Sets** | **Number of Items** |
| Magazine Promo = Yes | 7 |
| Watch Promo = No | 6 |
| Life Ins Promo = Yes | 5 |
| Life Ins Promo = No | 5 |
| Card Insurance = No | 8 |
| Sex = Male | 6 |

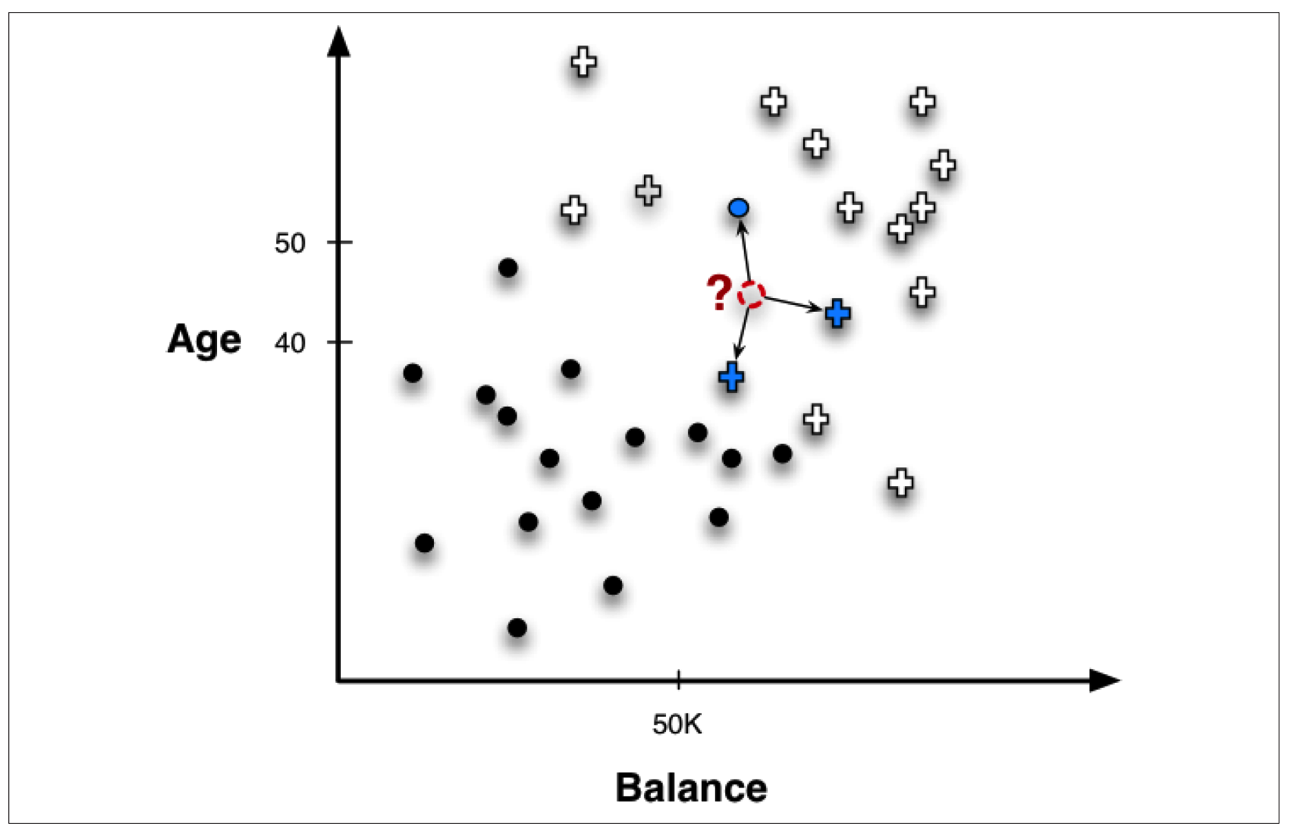
|  |  |
| --- | --- |
| **Two Item Sets** | **Number of Items** |
| Magazine Promo = Yes & Watch Promo = No | 4 |
| Magazine Promo = Yes & Life Ins Promo = Yes | 5 |
| Magazine Promo = Yes & Card Insurance = No | 5 |
| Watch Promo = No & Card Insurance = No | 5 |

1. One two-item set rule that can be generated from the tables above is: If Magazine Promo = Yes Then Life Ins promo = Yes. The confidence for this rule is:
2. **5 / 7**
3. 5 / 12
4. 7 / 12
5. 1
6. Based on the two-item set table, which of the following is not a possible two-item set rule?
7. IF Life Ins Promo = Yes THEN Magazine Promo = Yes
8. IF Watch Promo = No THEN Magazine Promo = Yes
9. IF Card Insurance = No THEN Magazine Promo = Yes
10. **IF Life Ins Promo = No THEN Card Insurance = No**
11. In estimating the accuracy of data mining classification models, true positive rate is:
12. **The ratio of correctly classified positives divided by the total positive count**
13. The ratio of correctly classified negatives divided by the total negatives count
14. The ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified posivitives
15. The ratio of correctly classified negatives divided by the sum of correctly classified positives and incorrectly classified negatives
16. Which of the following classifiers fall in the category of lazy learners?
17. Decision trees
18. Bayesian classifier
19. **K-NN classifier**
20. Rule-based classifier
21. In KNN classification, choosing high values of K results in:
22. **predicting the most frequent class label**
23. increasing the risk of overfitting due to noise in the training data
24. avoiding over-smoothing
25. increasing the risk of underfitting

**Problems/Mini Case Questions:**

**In the below diagram, there exist a point with a question mark “?”, which needs to be classified.**

* + - Explain which data mining technique should be used in this classification? k-Nearest neighbor classification [K-NN]
    - Explain how it could be used to label the “?”. The point to be classified, labeled with a question mark, would be classified + if we assume k=3, because the majority of its nearest (three) neighbors are +



Consider the below credit card marketing problem. The goal is to predict whether a new customer will respond to a credit card offer based on how other similar customers have responded. The data are shown in the below table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Customer** | **Age** | **Income (K)** | **Cards** | **Response (class label)** | **Distance from David** |
| **David** | **37** | **50** | **2** | **?** | **0** |
| **John** | **35** | **35** | **3** | **Y** | **15.6** |
| **Rachel** | **22** | **50** | **3** | **N** | **15** |
| **Ruth** | **63** | **200** | **1** | **N** | **152.23** |
| **Jefferson** | **59** | **170** | **1** | **N** | **122** |
| **Norah** | **25** | **40** | **4** | **Y** | **15.74** |

**Questions:**

1. Calculate the distance column.

*See table*

1. Given that K=3, what will be David’s response?

*Three are closely similar to David: Rachel, John, and Norah. Their responses are N, Y, Y respectively. If you take majority vote, then David will respond positively (Y).*

1. How does the answer change, if K=4?

*Yes, as we will have to include Jefferson and it is a N; i.e., then it is a tie 50%-50% then relative voting is required. {Relative majority is not required; no normalization is required}*

1. **A Classification:** Below is a two-class classification problem of predicting whether a photograph contains a man or a woman.



We have a test dataset of 10 records with expected outcomes and a set of predictions from our classification algorithm. See below table.

|  |  |  |
| --- | --- | --- |
|  | Expected | Predicted |
| 1 | Man | Woman |
| 2 | Man | Man |
| 3 | Woman | Woman |
| 4 | Man | Man |
| 5 | Woman | Man |
| 6 | Woman | Woman |
| 7 | Woman | Woman |
| 8 | Man | Man |
| 9 | Man | Woman |
| 10 | Woman | Woman |

* What is the classification accuracy for this set of predictions?

The algorithm made 7 of the 10 predictions correct with an accuracy of 70%.

* What is the error rate?

30%.

* Calculate the number of correct predictions for each class.

Men classified as men = 3

Woman classified as women = 4

* Calculate the number of incorrect predictions for each class, organized by the predicted value.

Men classified as women = 2

Woman classified as men = 1

* Arrange these values into the 2-class confusion matrix.

|  |  |  |
| --- | --- | --- |
|  | Men | Women |
| Men | 3 | 1 |
| Women | 2 | 4 |

* What is the total number of men in the dataset?

It is the sum of the values on the men column (3 + 2)

* What is the total number of women in the dataset?

It is the sum of values in the women column (1 +4)