1. Pete, owner of Pistol Pete’s Diamond Emporium, is investing in a diamond classification system due to his deteriorating eyesight. Pete buys and sells diamonds of varying quality: Low ($1,000-$3,000), Medium ($4,000-$7,000), and High ($8,000-$10,000). It is very important to Pete that his classifier properly classifies his diamonds so that he can not only have a profitable business, but also, so that his customers will continue to trust him as a business owner.

Using the *possible cost matrix values* given below, fill out the cost matrix that most accurately reflects Pete’s needs for his diamond classifier model. After completing the cost matrix, justify your proposed cost matrix.

*Possible cost matrix values*: -1, -1, 0, 20, 20, 20, 20, 100, 100

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Actual class** |  | **Predicted class** | | |
| High | Medium | Low |
| High |  |  |  |
| Medium |  |  |  |
| Low |  |  |  |

1. You have been given a data set containing three discrete attributes and five continuous attributes. After carefully analyzing the problem and the available attributes, you decide that one of the continuous attributes, *estimate*, should be used as the class attribute for your classification problem. Describe how you could use the *estimate* attribute when performing classification.
2. When performing an unsupervised k-means clustering, it is sufficient to generate a single clustering. Do you agree with this statement? Why or why not?
3. You have performed an unsupervised k-means clustering on a data set with two attributes and the results indicate a k value of 2. Later, a domain expert determines class values for each data instance and there is a total of four class values. Provide a possible explanation for why unsupervised clustering disagrees with the domain expert for a k value by drawing a sketch of the unsupervised clustering to go along with your explanation.
4. How many possible association rules can be generated from a transaction database containing 10 different items? If three of those items are infrequent, how many rules can be generated from all possible 2-itemsets if those 2-itemsets are all frequent? **Hint: recall how we used combinatorics to determine how many k-itemsets may be generated for a given number of items, k.**
5. Run the Nearest Neighbor classifier with a k-value of 7 and a Support Vector Machine with default values using 10-folds cross validation on the diabetes data set (diabetes.arff in Assignment 3 on myCourses) in Weka. Fill in the confusion matrices for the models in the tables below and use the cost matrix to compute the cost for each model. Based upon the **cost**, which model should be selected and why?

**Nearest Neighbor (k=7) Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Tested Negative** | **Tested Positive** |
| **Tested Negative** |  |  |
| **Tested Positive** |  |  |

**Support Vector Machine Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Tested Negative** | **Tested Positive** |
| **Tested Negative** |  |  |
| **Tested Positive** |  |  |

**Cost Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Tested Negative** | **Tested Positive** |
| **Tested Negative** | 0 | 50 |
| **Tested Positive** | 100 | -1 |