**Data Mining**

**Sample Midterm Questions (Last Modified 10/24)**

Please note that the purpose here is to give you an idea about the level of detail of the questions on the midterm exam. These sample questions are not meant to be exhaustive; you may certainly find topics on the midterm that are not covered here at all. Your midterm will include more questions than this.

* 1. Sometimes a data set is partitioned such that a validation set is provided. What is the purpose of the validation set?
1. Are decision trees easy to interpret (circle one): Yes No
2. How can you convert a decision tree into a rule set? Explain the process.
3. List two reasons why data mining is popular now and it wasn’t as popular 20 years ago.
4. How does an ordinal feature differ from a nominal feature? Explain in one or two sentences.
5. If we build a classifier and evaluate it on the training set and the test set:
	* 1. Which data set would we expect to have the higher accuracy: training set test set
		2. Which data set provides best accuracy estimate on new data: training set test set
6. For a two-class classification problem, with a Positive class P and a negative class N, we can describe the performance of the algorithm using the following terms: TP, FP, TN, and FN.
	* 1. What do each of these terms refer to?

TP:

TN:

FP:

FN:

* + 1. Place the 4 terms listed above in part a into the appropriate slots in the table below.

|  |  |  |
| --- | --- | --- |
|  |  | **Predicted** |
|  |  | **Positive** | **Negative** |
| **Actual** | **Positive** |  |  |
| **Negative** |  |  |

* + 1. Provide the formula for accuracy in terms of TP, TN, FP, and FN.
		2. Provide the formula for precision and recall using TP, TN, FP, and FN.

Precision =

Recall =

* + 1. What fraction of the total examples represented in the confusion matrix belong to the positive class?
		2. A learning curve shows the performance of a classifier as the *training set size* increases. Assume that training set size is plotted on the x-axis and accuracy is plotted on the y axis. On the figure below, plot a typical/expected learning curve when the accuracy is measured on the test set data.

|  |  |
| --- | --- |
| Accuracy  |   |

Training set size

1. You need to split on attribute *a1* in your decision tree. The attribute has 8 values. Why might a two-way split be better than an 8-way split? What might be a problem with the 8-way split?



1. Given a training set with 5+ and 10- examples,
	* 1. What is the entropy value associated with this data set? You need not simplify your answer to get a numerical answer.
		2. What is the Gini associated with this data set? In this case you should simplify your result, although you may express the answer the answer as a fraction rather than a decimal.
		3. If you generated a decision tree with just the root node for the examples in this data set, what class value would you assign and what would be the training-set error rate associated with this (very short) decision tree?
2. Dr. Weiss has stated that he believes that decision trees are more expressive than linear classifiers, that form a single linear decision boundary. Provide one reason why decision trees could be considered more expressive and one reason why one could argue linear classifiers are more expressive.
3. What is the curse of dimensionality?
4. Explain why accuracy is not an appropriate evaluation metric when the classes are highly imbalanced?
5. Very often we utilize F-measure rather than precision and recall. What is the advantage of using only a single metric?
6. Fill in the following for the Area Under the ROC curve (AUC):

Minimum AUC value:

Maximum AUC value:

AUC value if guessing:

1. If you utilize a decision tree algorithm that employs pruning and then you double the number of training examples, do you expect the number of nodes in the generated tree to decrease, stay the same, or increase?
2. One method to deal with class imbalance is to oversample the minority class by making exact copies. What is the drawback of this method?
3. List as many positives/advantages of decision trees as you can. A minimum of three is required.
4. There will be a decision tree splitting problem similar to the ones on HW2. Make sure you understand Gini, Entropy, and classification error rate. Redo some of the questions on HW2.