Data Mining
Sample Midterm Questions (Last Modified 2/17/19)

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 and 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?

2. Are decision trees easy to interpret (circle one): Yes No

3. How can you convert a decision tree into a rule set? Explain the process.

4. List two reasons why data mining is popular now and it wasn’t as popular 20 years ago.

5. How does an ordinal feature differ from a nominal feature? Explain in one or two sentences.

6. Sally measures the pressure of all of tires coming into her garage for an oil change and records the values. Unknown to her, her tire gauge is miscalibrated and adds 3 psi to each reading. According to the definition of noise used by our textbook, is this error introduced by the tire gauge considered noise? Answer “yes” or “no” and justify your answer.

7. 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.

   a) What do each of these terms refer to?

      TP:

      TN:

      FP:

      FN:

   b) Place the 4 terms listed above in part a into the appropriate slots in the table below.

		Predicted	
		Positive	**Negative**
Actual	**Positive**		
	Negative		

   c) Provide the formula for accuracy in terms of TP, TN, FP, and FN.

   d) Provide the formula for precision and recall using TP, TN, FP, and FN.

      Precision =

      Recall =

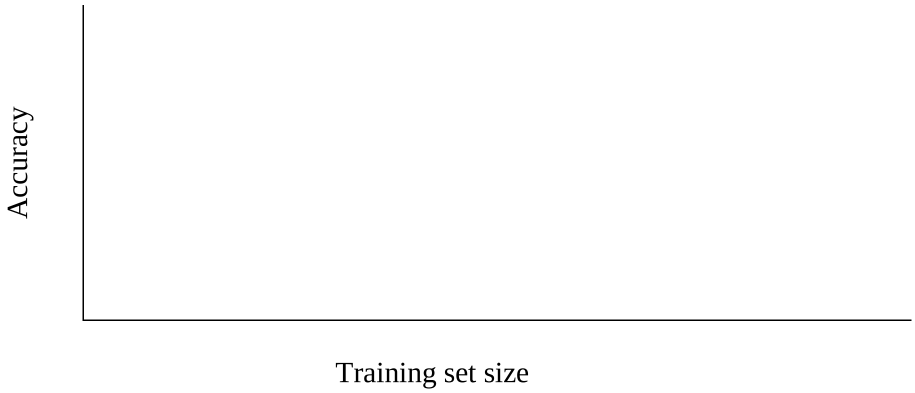
8. If we build a classifier and evaluate it on the training set and the test set:

   a) Which data set would we expect to have the higher accuracy: training set test set

   b) Which data set provides best accuracy estimate on new data: training set test set

9. 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.

   a) On the figure below, plot a typical/expected learning curve when the accuracy is measured on the 1) training set data and 2) the test set data (i.e., draw two curves). Should there be any difference? If so, comment on the expected difference.



10. 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?

$$Entropy(t) = -\sum_j p(j\,|\,t) \log p(j\,|\,t)$$

$$GINI(t) = 1 - \sum_j [p(j\,|\,t)]^2$$

11. Given a training set with 5+ and 10- examples,

   a) What is the entropy value associated with this data set? You need not simplify your answer to get a numerical answer.

   b) 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.

c) 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?

12. The nearest neighbor algorithms relies on having a good notion of similarity, or distance. In class we discussed several factors that can make it non-trivial to have a good similarity metric. What were two of the factors?

13. What classifier induction algorithm can effectively generate the most expressive classifiers, in terms of the decision boundaries that can be formed? Which is the least expressive. Rank order them from most to least expressive. Briefly justify your ordering.

    The induction algorithms are: decision trees, linear classifiers, and nearest neighbor.

14. What is the curse of dimensionality?

15. What does it mean if the rule set for a rule learner is exhaustive?

16. Does the Ripper Rule Learner build rules from general to specific or specific to general?