Data Mining
Sample Midterm Solutions (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?

   The validation set is used to select amongst multiple models or to tune a specific model (which can be viewed as a family of models).

   Here is more explanation. In this sense is it used like a test set in that it is used for evaluation, but it is not like a test set in that it cannot be used for reporting the performance of the model. That is, the validation set chooses a model—for example the right amount or pruning—but then that model can be evaluated on a test set and that performance number can be reported. If one dataset could be used for finding the best model and reporting the performance, you could generate 1million models and then pick the best on the data set and report that performance. But it is likely that model really does not have the best performance but just happened to do best on that one data 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.

   Create one rule per leaf node by traversing the conditions from the root node to the leaf and conjoining those conditions. Note that the rules would be mutually exclusive, meaning that only one rule could "fire" at a time.

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

   Faster computers, cheaper memory, more data being routinely recorded (e.g., popularity of the Web and devices like smartphones), and to a lesser degree better algorithms.

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

   An ordinal feature is a nominal feature where there is a natural ordering of each attribute value.

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.

   No, since noise must be random, not systematic.

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: *True Positive*

   TN: *True Negative*

   FP: *False Positive*

   FN: *False Negative*

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

		Predicted	
		Positive	**Negative**
Actual	**Positive**	*TP*	*FN*
	Negative	*FP*	*TN*

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

   (TP + TN) / (TP + TN + FP + FN)

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

   Precision = *TP/(TP + FP)*

   Recall = *TP/(TP + FN)*

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.

   The most important thing is that for the test set data it increases, initially steeply but then begins to plateau. Do not worry too much about the training set curve, since the answer is not really clear.

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?

    The 8-way split can lead to the problem of data fragmentation. The data will be split up excessively leaving smaller amounts of data available for future splits.

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t) \qquad GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

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.

$$-(1/3)\log(1/3) - (2/3)\log(2/3)$$

    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.

$$1 - [(1/3)^2 + (2/3)^2] = 1 - 1/9 - 4/9 = 1 - 5/9 = 4/9$$

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?

Majority class is negative class, so classify it as negative. Training error rate is then 5/15 1/3.

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?

A good similarity metric requires that the scales of the features are similar. For example, if one feature varies from 1 to 100 and another from 1, to 1,000,000, then there is a problem and the values should be rescaled. Another problem is that some features may be much less important than others, and yet by default all features are considered equally important. Also, redundant or highly correlated features will through off the distance metric, because the related features will be overvalued.

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.

Most expressive: nearest neighbor

Middle: decision trees

Least expressive: linear classifier

14. What is the curse of dimensionality?

The curse of dimensionality is that when the number of features increases, the concentration of the data points within the instance space decreases, which makes it harder to find patterns. For example, if you have 100 data points and one variable, it is likely the space is dense, but if we have 100 features, the space will be quite sparse.

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

It means that the rules will collectively cover every possible example.

16. Does the Ripper rule learner build rules from general to specific or specific to general?

It builds rules from general to specific. It starts with a rule where the antecedent has no conditions and then adds conditions one at a time.