# Data Mining
# Sample Midterm Solutions (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?

   *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. 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]

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 part a into 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)*

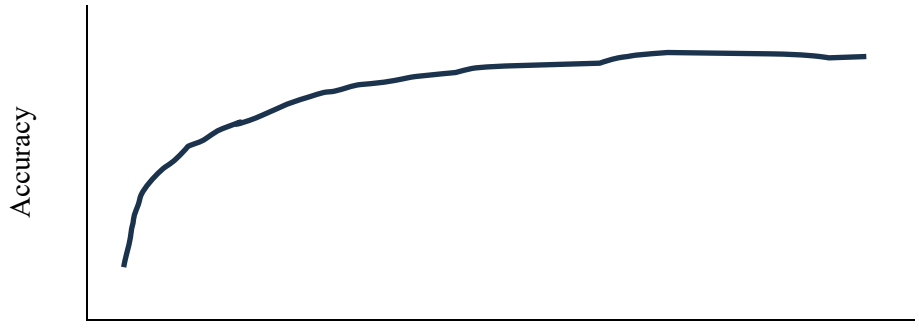
Recall = *TP/(TP + FN)*

e) What fraction of the total examples represented in the confusion matrix belong to the positive class?

*The number of positive examples is TP + FN so the fraction is:*

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

f) 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.




*The most important thing is that it initially increases steeply but then begins to plateau.*

8. 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.*

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

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

9. 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, so classify it as negative. Training error rate is then 5/15= 1/3.*

10. 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.

    *Decision trees are more expressive because they can generate many decision boundaries (i.e., rectangles) so they can partition the space into many pieces, while a linear classifier can only form one boundary and break things into two pieces. However, a linear classifier could be considered more expressive in one way: it can learn non-axis parallel (i.e, slanted) boundaries, while a decision tree can only learn axis-parallel decision boundaries.*

11. 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.*

12. Explain why accuracy is not an appropriate evaluation metric when the classes are highly imbalanced?

    Accuracy effectively weights the performance of the classes values based on the fraction of the training data that they represent, so if one class occurs 5 more times than another, it will have five times the impact on accuracy. Thus accuracy does not count the performance of the minority class enough and this may lead to strategies that always, or almost always, predict the majority class.

13. Very often we utilize F-measure rather than precision and recall. What is the advantage of using only a single metric?

    It makes it easier to compare performance as with two values you may have one algorithm/model that does better on one metric but worse on the other and in such cases it is not clear which one is better.

14. Fill in the following for the Area Under the ROC curve (AUC):

    Minimum AUC value: 0

    Maximum AUC value: 1

    AUC value if guessing: 0.5

15. 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?

    Increase as with more data there is more to learn. Pruning will prevent overfitting, but in most cases with more data you can learn more details. It might remain the same if your model has already plateaued, but that generally does not happen.

16. One method to deal with class imbalance is to oversample the minority class by making exact copies. What is the drawback of this method?

    By making exact copies the classifier will be more likely to overfit those examples that are duplicated multiple times. Overfitting.

17. List as many positives/advantages of decision trees as you can. A minimum of three is required.

    Easily explained and comprehended; very quick to build the model; very quick to apply the model to test cases; automatically handles redundant and irrelevant features

18. 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.