Problem Set 3

CS 4375

Due: 11/2/2025 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. **NO** ML libraries are permitted. Late homeworks will not be accepted.

Problem 1: Poisonous Mushrooms? (20 pts)

For this problem, you will use the mushroom data set provided with this problem set. The data has been divided into two pieces mush_train.data and mush_test.data. These data sets were generated using the UCI Mushroom data set (follow the link for information about the format of the data). Note that the class label is the first column in the data set.

1. Assuming you break ties using the attribute that occurs **first** (left to right) in the data, draw the resulting decision tree and report the maximum information gain for each node that you added to the tree.

2. What is the accuracy of this decision tree on the test data?

Problem 2: Cross-Validation (20 pts)

Using a single tuning set for the hyperparameters can yield an unreliable predictor of the class label, i.e., maybe it was not a representative sample of the data, plus some data is "wasted" using this approach. An alternative approach that is particularly applicable for small data sets is k-fold cross-validation.

1. Partition the non-test data into k equally sized buckets.

2. For each possible set of hyperparameters you will train the model using exactly $k-1$ of the partitions while the held out partition is used as a validation data set.

3. As there are k different ways to hold out one partition, all k possibilities are tried and the average validation set accuracy (as measured by the appropriate held-out data) of the k different models learned for each of the hyperparameter settings is used to select the winning hyperparameters.

4. Finally, the model is retrained using all of the non-test data with the winning hyperparameters and then evaluated using the test data.

Apply 10-fold cross validation to fit an SVM with slack classifier (no feature maps) to the data set wdbc_train.data (each row corresponds to a single data observation and the class label +1/-1 is the first entry in each row). The partitions for cross validation should be selected as equally sized contiguous blocks of data starting from the first data element. You should choose the ranges for the hyperparameters (make sure the range is reasonable). Report the best setting of the hyperparameters and the accuracy on the test set wdbc_test.data.

Problem 3: Medical Diagnostics (60 pts)

For this problem, you will use the data set provided with this problem set. The data has been divided into two pieces heart_train.data and heart_test.data. These data sets were generated using the UCI SPECT heart data set (follow the link for information about the format of the data). Note that the class label is the first column in the data set.

1. Suppose that the hypothesis space consists of all decision trees with exactly one attribute split for this data set.

   (a) Run the adaBoost algorithm for 10 rounds to train a classifier for this data set. Draw the 10 selected trees in the order that they occur and report the ϵ and α, generated by adaBoost, for each.

   (b) Plot the accuracy on the training and test sets versus iteration number.

   (c) Use coordinate descent to minimize the exponential loss function for this hypothesis space over the training set. You can use any initialization and iteration order that you would like other than the one selected by adaBoost. What is the optimal value of α that you arrived at? What is the corresponding value of the exponential loss on the training set?

   (d) Use bagging, with 20 bootstrap samples, to produce an average classifier for this data set. How does it compare to the previous classifiers in terms of accuracy on the test set?

   (e) Which of these 3 methods should be preferred for this data set and why?