Problem Set 2

CS 6375

Due: 9/30/2017 by 11:59pm

Note: all answers should be accompanied by explanations and relevant code for full credit. Late homeworks will not be accepted.

Problem 1: Breast Cancer Diagnosis (50 pts)

For this problem, you will use the cancer data set provided with this problem set. The data has been divided into three pieces wdbc_train.data, wdbc_validation.data, and wdbc_test.data. These data sets were generated using the UCI Breast Cancer Wisconsin (Diagnostic) data set (follow the link for information about the format of the data). Note that the class label (malignant or benign) is the first column in the data set. All code (Python or MATLAB only) should be turned in with your answers to the following questions.

1. Primal SVMs
   (a) Use the SVM with slack formulation to train a classifier for each choice of $c \in \{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8\}$ without using any feature maps.
   (b) What is the accuracy of the learned classifier on the training set for each value of c?
   (c) Use the validation set to select the best value of c. What is the accuracy on the validation set for each value of c?
   (d) Report the accuracy on the test set for the selected classifier.
2. Dual SVMs with Gaussian Kernels
   (a) Use the dual of the SVM with slack formulation to train a classifier for each choice of $c \in \{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8\}$ using a Gaussian kernel with $\sigma \in \{.1, 1, 10, 100, 1000\}$.
   (b) What is the accuracy of the learned classifier on the training set for each pair of c and σ?
   (c) Use the validation set to select the best value of c and σ. What is the accuracy on the validation set for each pair of c and σ?
   (d) Report the accuracy on the test set for the selected classifier.
3. What is the accuracy of the k-nearest neighbor classifier for $k = 1, 5, 11, 15, 21$? You don't need to implement the k-dimensional tree version.
4. Which of these approaches (if any) should be preferred for this classification task? Explain.

Problem 2: Poisonous Mushrooms? (50 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. Train a decision tree using the information gain heuristic to select attributes as described in class (break ties using the attribute that occurs last (left to right) in the data). Draw it.
2. What is the size (number of nodes) in the learned decision tree?
3. What is the height of the learned decision tree?
4. What is the accuracy of your learned decision tree on the training set?
5. What is the accuracy of your learned decision tree on the test set?
6. The Audubon Society Field Guide to North American Mushrooms states that there is not a simple set of rules to determine whether or not a mushroom is edible. How well would you say that decision tree learning works for this problem?
7. How dependent is the quality of the learned decision tree on the training/test split? Explain.
8. Is the best decision tree with exactly one non-leaf node for this training set equal to the one found by using the greedy heuristic to select one attribute?