Problem Set 2

CS 6375

Due: 9/28/2015 by 11:59pm

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

Problem 1: Spam, Spam, Ham (45 pts)

For this problem, you will use the spam data set provided with this problem set. The data has been divided into three pieces spam_train.data, spam_validation.data, and spam_test.data. These data sets were generated using the UCI Spambase data set (follow the link for information about the format of the data). Note that the class label is the last column in the data set.

1. Primal SVMs
   (a) Use the SVM with slack formulation to train a classifier for each choice of $c \in \{1, 10, 100, 1000, 10000\}$.
   (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 your accuracy on the test set.
2. Dual SVMs with Gaussian Kernels
   (a) Use the dual of the SVM with slack formulation to train a classifier for each choice of $c \in \{1, 10, 100, 1000, 10000\}$ using a Gaussian kernel with $\sigma \in \{.001, .01, .1, 1, 10, 100\}$.
   (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 your accuracy on the test set.
3. Which of these approaches (if any) should be preferred for the spam email classification task? Explain.

Problem 2: Poisonous Mushrooms? (45 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 a majority vote).
2. What is the size (number of nodes) in the learned decision tree?
3. What is the depth 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)?

Problem 3: Understanding Nearest Neighbor Methods (10 pts)

Let's suppose that we have two distinct training sets S_1 and S_2 where the labels are either $+$ or $-$.

1. Suppose that you are given a new data point x to classify. Argue that if the 1-nearest neighbor algorithm labels x as $+$ using S_1 as a training set and if the 1-nearest neighbor algorithm labels x as $+$ using S_2 as the training set, then the 1-nearest neighbor algorithm will label x with a $+$ when using $S_1 \cup S_2$ as the training set.
2. Provide an example to illustrate that this is not true for the 3-nearest neighbor algorithm.