Problem Set 2

CS 6375

Due: 3/10/2020 by 11:59pm

Note: all answers should be accompanied by explanations and relevant code for full credit. All code (Python or MATLAB only) should be turned in with your answers to the following questions. Late homeworks will not be accepted.

Problem 1: Parkinson's Disease (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 park_train.data, park_validation.data, and park_test.data. These data sets were generated using the UCI Parkinsons Data Set data set (follow the link for information about the format of the data). Note that class label, health status of the subject, 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) Using gradient descent or quadratic programming, apply 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) Using quadratic programming, apply 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 $\sigma^2 \in \{.1, 1, 10, 100, 1000\}$.
   (b) What is the accuracy of the learned classifier on the training set for each pair of c and σ^2?
   (c) Use the validation set to select the best value of c and σ^2. What is the accuracy on the validation set for each pair of c and σ^2?
   (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: Method of Lagrange Multipliers (20 pts)

Suppose that we modified the objective function in the SVM with slack formulation to be a quadratic penalty instead of a linear penalty, that is minimize $\frac{1}{2}||w||^2 + c\sum_i \xi_i^2$ subject to the same constraints as the standard SVM with slack. What is the dual of this new quadratic penalized SVM with slack problem for a fixed c? Can the kernel trick still be applied?

Problem 3: Poisonous Mushrooms? (30 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 **last** (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?

3. Now consider arbitrary input data. Suppose that you decide to limit yourself to decision trees of height one, i.e., only one split. Is the tree produced by the information gain heuristic optimal on the training data (that is, no other decision tree has higher accuracy)?