Problem Set 3

CS 4375

Due: 10/18/2022 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: VC Dimension (35 pts)

1. Given training data of the form $(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$, where $x^{(m)} \in \mathbb{R}^n$ and $y^{(m)} \in \{-1, 1\}$, consider the hypothesis space of n-dimensional spheres: each element of the hypothesis space is parameterized by a center $c \in \mathbb{R}^n$ and a radius $r > 0$ such that all points within distance r of the center c are classified as $+1$ and the remaining points are classified with a -1. What is the VC dimension of this hypothesis space?

2. Consider a binary classification problem for data points in $\mathbb{R}^2$ with a hypothesis space consisting of pairs of parallel lines such that any point between the pair is classified as $'+'$ and points outside of the pair are classified as $-$. What is the VC dimension of this hypothesis space? Prove it. How many samples would be sufficient to guarantee that an optimal learning algorithm will attain an accuracy of .8 with probability at least .95?

Problem 2: Medical Diagnostics (65 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?