

Problem Set 3

CS 6375

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

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

Problem 1: VC Dimension (35 pts)

1. 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?
2. Consider a binary classification problem for data points in \mathbb{R} . Let H be the hypothesis space of all intervals in \mathbb{R} . Given an interval in H , points inside the interval are classified as '+' and the remaining points are classified as '-'. Consider the boosted hypothesis space H' that takes a pair of hypotheses from H and takes the sign of their weighted combination (similar to what would be produced by two rounds of boosting). Specifically,

$$H' = \{f | f(x) = \text{sign}(\alpha_1 h_1(x) + \alpha_2 h_2(x)) \text{ for some } h_1, h_2 \in H \text{ and } \alpha_1, \alpha_2 \in \mathbb{R}\}.$$

To break ties, if $\alpha_1 h_1(x) + \alpha_2 h_2(x) = 0$, the hypothesis should return a '+'. What is $VC(H')$? Prove it.

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 two attribute splits (repetition along the same path is allowed) for this data set.
 - (a) Run the adaBoost algorithm for five rounds to train a classifier for this data set. Draw the 5 selected trees in the order that they occur and report the ϵ and α , generated by adaBoost, for each.
 - (b) Run the adaBoost algorithm for 10 rounds of boosting. Plot the accuracy on the training and test sets versus iteration number.

2. Now, suppose that the hypothesis space consists of only height 1 decision trees for this data set (only one attribute split).
- (a) 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?
 - (b) What is the accuracy of the resulting classifier on the test data?
 - (c) What is the accuracy of adaBoost after 20 rounds for this hypothesis space on the test data? How does the α learned by adaBoost compare to the one learned by gradient descent?
 - (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?