10-601 Machine Learning: Homework Assignment 5 Problem 1 Solution

1 Computational Learning Theory

1.1 VC dimension

Consider the space of instances X corresponding to all points in the x, y plane. Give the VC dimension of the following hypothesis spaces:

1. H_r = the set of all rectangles in the x, y plane. That is, $H = \{((a < x < b) \wedge (c < y < d))|a, b, c, d \in \mathbb{R}\}$.

   ⋆ *Solution*: $VC(H) = 4$ if you interpreted that points inside the rectangle are always classified as positive examples, and $VC(H) = 5$ if you interpreted that you were allowed to choose whether points inside or outside were classified as positive examples.

2. H_c = circles in the x, y plane. Points inside the circle are classified as positive examples.

   ⋆ *Solution*: $VC(H) = 3$

3. H_t = triangles in the x, y plane. Points inside the triangle are classified as positive examples.

   ⋆ *Solution*: $VC(H) = 7$

1.2 Probably approximately correct (PAC) learning

Consider a decision tree learning algorithm that considers only examples described by Boolean features $\langle X_1, \ldots X_n \rangle$, learns only Boolean-valued functions ($Y \in \{+, -\}$), and outputs only 'regular, depth-2 decision trees.' A 'regular, depth-2 decision tree' is a depth two decision tree (a tree with four leaves) in which the left and right child of the root are *required to test the same attribute.* For example, the tree in Figure 1 is a 'regular, depth-2 decision tree.'

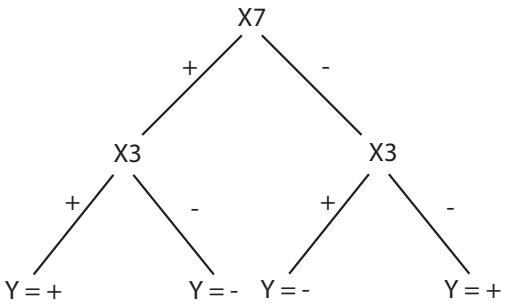

Figure 1: An example of a regular Boolean depth-2 decision tree.

1. Suppose you have noise-free training data for target concept c which you know can be described by a regular, depth-2 decision tree. How many training examples must you provide the learning algorithm in order to assure that with probability .99 the learner will output a tree whose true accuracy is at least .97? Assume you have data with 20 attributes in total

(though of course you believe only two of these twenty will be needed to describe the correct tree).

⋆ *Solution*: Since $|H|$ is finite, we can use the bound $m \geq \frac{1}{\epsilon}(ln(1/\delta) + ln|H|)$. The number of unique hypotheses is $20 * 19 * 2^4 = 6080$. One could plug this in and get a loose bound. However, if we consider all of the hypotheses that are equivalent (many hypotheses provide the same mapping from the instance space to the output space— for example, the ordering of the two attributes that are chosen does not matter), we can find a tighter bound. There are 2 hypotheses that label all examples with one label (either all positive or negative).

There are $20 * 2$ hypotheses that split on one attribute, and label one side positive, and the other negative.

There are $\binom{20}{2} * 10$ different hypotheses that split on two attributes and assign labels to the leaves that are not equivalent to the hypotheses mentioned previously (to see this, consider that there are $2^4 = 16$ ways to assign labels to the leaves given some choice of attributes to split on, but the cases {++++, - - - -, ++ - -, - - ++, + - + -, - + - +} have already been covered).

Thus, $|H| = 2 + 20 * 2 + \binom{20}{2} * 10 = 1942$.

So our bound is: $m \geq \frac{1}{.03}(\ln(1/.01) + \ln(1942)) = 405.888$.

So 406 examples are required.

Thanks to Lawrence Jesper for his elegant solution to finding the number of unique hypotheses.

2. Suppose you modify the algorithm to handle instances that have real-valued attributes instead of Boolean attributes, and you allow each decision tree node to perform a Boolean threshold test of the form $X_i > a$ where a is allowed to take on any real value. The tree is further constrained such that the two second level nodes, must both test the same attribute *and* use the same threshold. In this case, re-answer the above question: How many training examples must you provide the learning algorithm in order to assure that with probability .99 the learner will output a tree whose true accuracy is at least .97? In this case, assume that each example has only two attributes, so the tree will end up using both. You can still assume that the target concept c is in the new hypothesis space.

⋆ *Solution*: Because of the real-valued thresholds that we choose in our hypotheses, the size of the hypothesis space is infinite. Thus, we use a VC dimension-based bound in this case. You should be able to convince yourself that the VC dimension is 4.

Using this:

$$m \geq \frac{1}{.03}(4\log_2(2/.01) + 8 * 4 * \log_2(13/.03)) = 10362.47$$

So, we need to provide at least 10363 examples.

Some students also provided a lower bound on the sample complexity, using the result from Ehrenfeucht et al. provided in class. This was not required to answer the question posed in the problem.