Assignment 4

CSG220, Spring 2007
Due: Thursday, March 15

1. Consider training a two-input perceptron on a set of training data of size R, where the data is assumed to be drawn i.i.d. from an unknown distribution.

(a) Suppose it is known that the true classification of all possible data is realizable by a linear separator (and the classification of the training data is noise-free). Use the appropriate PAC-learning bound from the textbook to compute a worst-case upper bound on the number of training examples R sufficient to assure with 90% confidence that a perceptron trained to zero error will have true error rate at most 5%.

(b) Now suppose it is not known whether the true classification of all possible data is realizable by a linear separator. Suppose that applying some training algorithm to 2000 data points gives a linear separator with 3% error on this training set. Use the appropriate error bound from the lecture notes to give a 95% confidence interval within which the true error rate of this classifier lies.

2. The purpose of this problem is to have you perform a thorough first-principles analysis of a trivially simple learning scenario, under some varying assumptions that include the PAC-learning assumptions. Here is the learning task: There are two instances, 0 and 1, and there are two possible classifications, which we also call 0 and 1. This problem can be restated as the problem of learning a Boolean function of a single Boolean variable. We assume that training examples are noise-free. The learner we consider simply stores a table associating each instance seen with its class. When asked to predict the class of any test instance it either produces the stored class if there is one or flips a fair coin to determine its answer otherwise.

For all answers you give below, provide a rigorous mathematical justification.

a. In the *query model of learning*, it is assumed that every training instance is selected by the learner, and the "teacher" supplies the correct classification for whatever instance it is queried about. (i) What is the minimum number of queries necessary to guarantee that the learner makes no errors on any subsequent test instance? (ii) Now suppose that test instances are drawn uniformly randomly from $\{0, 1\}$. What would be the minimum number of queries required if the learner is allowed to have a probability of 0.5 of committing an error on a subsequent test instance? (iii) What is the minimum number of training examples (i.e., queries) needed to insure with probability at least 90% that the learner's error probability on a test example is ≤ 0.05 (assuming once again that test instances are drawn uniformly randomly from $\{0, 1\}$)?

In the remaining parts of this problem we focus on *passive learners*, for whom training examples are generated by some distribution over the instance space from which correctly classified instances are drawn randomly. A passive learner has no control over which instances it sees during training. In the learning scenario we consider here, there are only 2 possible instances, so the underlying distribution can be characterized as being governed by a probability $p \in [0, 1]$ of drawing the instance 1. Both the training and test examples are drawn i.i.d. from this distribution.

b. In the *mistake bound* measure of learning performance, we want to determine the maximum number of errors the learner might make before never making another one. Give a tight mistake bound for this learner in this learning scenario.

Now we turn to a PAC-learning analysis of this specific learning problem. Introduce the notation $MinSampleSize(\delta, \epsilon)$ to denote the minimum number of training examples necessary to guarantee with probability at least $1 - \delta$ that the true error rate of the learner (after having seen this many training examples) is no larger than ϵ. Recall that the *true error rate* is defined to be the expected error on a subsequent randomly drawn test instance, where this expectation is taken over the true (but unknown to the learner) distribution from which training and test examples are drawn. Also, remember that in the PAC-learning analysis we want a guarantee that holds for *any* distribution (i.e., for any p in this two-instance learning problem).

c. What is the size of the hypothesis space for this learning problem? Use this in the appropriate formula from the textbook to compute an upper bound on $MinSampleSize(\delta, \epsilon)$, for all combinations of $\delta = 0.1, 0.05, 0.01$ and $\epsilon = 0.0, 0.1, 0.05, 0.01$. Organize your results into a table for easy inspection. (Some entries may be infinite.)

We now determine *exact* values for $MinSampleSize(\delta, \epsilon)$ as a function of these same 12 combinations of δ and ϵ in this simple learning problem, to compare against the bounds of part c, under various assumptions on what p can be.

d. Suppose $R \geq 1$ training examples have been seen (randomly generated i.i.d. according to the unknown p). How can it happen that the learner can have nonzero true error after seeing this many training examples? For each such scenario in which this might happen, determine the true error rate for this scenario and also determine its probability of occurrence. Summarize this information as a probability mass function on the true error rate as a function of p. (Hint: There are two such nonzero-error-rate scenarios involving $R \geq 1$ training examples. When $p = 0.5$, they give rise to the same error rate, and when $p \neq 0.5$, they give rise to different error rates. Thus, counting the zero-error-rate case, this probability mass function assigns nonzero probability to two points if $p = 0.5$ and three points if $p \neq 0.5$.)

Note: The largest value of ϵ for which $MinSampleSize(\delta, \epsilon)$ is to be tabulated is 0.1. Therefore, you should assume throughout all subsequent analyses that $\epsilon < 0.25$. (If this assumption is not made, there could potentially be other sets of answers applicable to cases like $\epsilon = 0.5$, etc.)

e. Suppose we know that $p = 0.5$. Derive an exact formula for the probability that the true error rate is greater than ϵ for any $\epsilon < 0.25$, as a function of R. Use this to determine an explicit formula for $MinSampleSize(\delta, \epsilon)$ as a function of $\delta > 0$ and ϵ that applies to this particular $p = 0.5$ case. Remember to apply the appropriate floor or ceiling function to guarantee that this is an integer. Then tabulate values of this function for the 12 combinations of values $\delta = 0.1, 0.05, 0.01$ and $\epsilon = 0.0, 0.1, 0.05, 0.01$. (Hint: In this case, you should find that none of these $MinSampleSize(\delta, \epsilon)$ values is infinite.)

For the remaining parts of this problem, assume that $p \neq 0.5$. Since there is a complete symmetry between the $p < 0.5$ and $p > 0.5$ cases, for convenience you should further restrict attention to the $p < 0.5$ case.

f. Under the assumption that $p < 0.5$, give a formula for the probability that the true error rate is greater than ϵ as a function of p and R. You should find that there are two cases (under the assumptions that $\epsilon < 0.25$ and $p < 0.5$), depending on the relative sizes of the user-selectable (hence arbitrary) ϵ and the fixed (but unknown) p.

g. Now assume a given value of ϵ (< 0.25). Apply the results of part f to determine what value of $p < 0.5$ (expressed a function of ϵ) maximizes the probability that the true error rate is greater than ϵ, and give an explicit formula for the maximum value of this probability. It will be a function of ϵ and R. Hints: (1) You may take for granted the easily proven fact that the function $f(x) = x^n + (1 - x)^n$ is

nonincreasing over [0,0.5] for $n \geq 1$. (2) You should find that even though there are two cases in part f, one of them always dominates when it comes to maximizing the probability in question for $p \in [0, 0.5)$.

h. Use the formula of part g to determine $MinSampleSize(\delta, \epsilon)$ for the 12 combinations of values $\delta = 0.1, 0.05, 0.01$ and $\epsilon = 0.0, 0.1, 0.05, 0.01$. Since you will not be able to invert the formula of part g to obtain an explicit formula for the value of R corresponding to a given δ, you should instead tabulate the formula of part g for appropriate fixed ϵ values and appropriate integer values of R and then determine $MinSampleSize(\delta, \epsilon)$ for the appropriate combinations of δ and ϵ values by inspection. (Using Excel is one handy way to do this.) As before, list these values in a table. If any entries in your table are infinite, provide proofs that the corresponding quantities are unbounded.

i. [Extra Credit] The standard PAC-learning analysis is based on the assumption that the distribution over the instance space (used in both training and testing) is completely unknown, and the computations are designed to allow for the possibility of a worst-case distribution. Thus the $MinSampleSize$ function you have computed in part h is designed to be valid for any possible p. Note that in part e we have already considered one situation where some prior knowledge of p (namely, $p = 0.5$) avoids worst-case behavior and gives considerably lower $MinSampleSize(\delta, \epsilon)$ values.

We now consider an intermediate situation, in which p is not known exactly, but we do know that it is bounded away from its most extreme values of 0 or 1 by a given amount. In particular, suppose that we know that $p \in [0.1, 0.9]$. Compute exact values of $MinSampleSize(\delta, \epsilon)$ for the same 12 combinations of values $\delta = 0.1, 0.05, 0.01$ and $\epsilon = 0.0, 0.1, 0.05, 0.01$ under this assumption.

3. Recall the "circle machine" discussed in lecture that classifies points $\mathbf{x}$ in the plane according to

$$f(\mathbf{x}, b) = sgn(\mathbf{x} \cdot \mathbf{x} - b),$$

where b is an arbitrary scalar.

a. Define a "hypersphere machine" to be the generalization of the circle machine to a classifier that classifies points $\mathbf{x}$ in m-dimensional space according to the prescription

$$f(\mathbf{x}, b) = sgn(\mathbf{x} \cdot \mathbf{x} - b),$$

where b is an arbitrary scalar. Derive its VC-dimension, and prove your answer.

b. Now define a generalization of this that allows the center of the hypersphere to be located anywhere, not just the origin. That is, this new machine classifies m-dimensional points $\mathbf{x}$ according to

$$f(\mathbf{x}, \mathbf{c}, b) = sgn((\mathbf{x} - \mathbf{c}) \cdot (\mathbf{x} - \mathbf{c}) - b),$$

where $\mathbf{c}$ is any m-dimensional point and b is any real number. Prove that the VC-dimension of this machine is at least $m + 1$. (Hint: Try to take advantage of other known VC-dimension results.)