CMPS 142 Fourth Homework, Spring 2015

2 Problems, 7 pts, due start of class Tuesday May 19

This homework is to be done in groups of 2 or 3. Each group members should completely understand the group's solutions and *must* acknowledge all sources of inspiration, techniques, and/or helpful ideas (web, people, books, etc.) other than the instructor, TA, and class text. Each group should submit a single set of solutions containing the names and e-mail addresses of all group members. Although there are no points for "neatness", the TA may deduct points for illegible or poorly organized solutions.

1. (4 pts) This problem guides you through a simple PAC-style (Probably Approximately Correct) generalization bound for a boolean classification problem. Let the domain be the unit interval in one dimension, i.e. the closed interval $[0, 1]$. Any *initial segment* of the unit interval is associated with a threshold T and consists of those points in the (closed) interval $[0, T]$. Let the hypothesis class $\mathcal{H}$ be the set of initial segments, so $\mathcal{H} = \{h_z : 0 \leq z < 1\}$. where each $h_z \in \mathcal{H}$ predicts $+$ on those points in $[0, z]$, and $-$ on those points in the half-open interval $(z, 1]$.

   Assume that the target concept is some h_θ (which is unknown to the algorithm). Assume that there is a continuous probability density p over $[0, 1]$ and each example (x, y) is generated independently by first sampling from p to get the feature value x and then setting the label y to be $+$ if $x \leq \theta$ and $y = -$ if $x > \theta$. Thus there is no noise in the data. Consider algorithm A that outputs the most specific consistent hypothesis from an m-example training set $D = \{(x^{(1)}, y^{(1)}), \ldots (x^{(m)}, y^{(m)})\}$. In other words, A outputs $h_{\hat{\theta}}$ where $\hat{\theta} = \max_{1 \leq i \leq m}\{x^{(i)} : y^{(i)} = +\}$ (the maximum of those points in D that are labeled $+$). Since the data is assumed noise-free, $\hat{\theta} \leq \theta$.

   We first notice that any h_z with $z < \theta$ predicts incorrectly on a new random example (x, y) exactly when x falls in the half-open interval $(z, \theta]$. Therefore the error rate of $h_{\hat{\theta}}$ is the probability under density p (which is unknown to the algorithm) of the interval $(\hat{\theta}, \theta]$.

   Now consider an arbitrary error tolerance ϵ and let z_ϵ be the value such that the interval $p((z_\epsilon, \theta]) = \epsilon$ (assume that $p((0, \theta]) > \epsilon$ so z_ϵ exists).

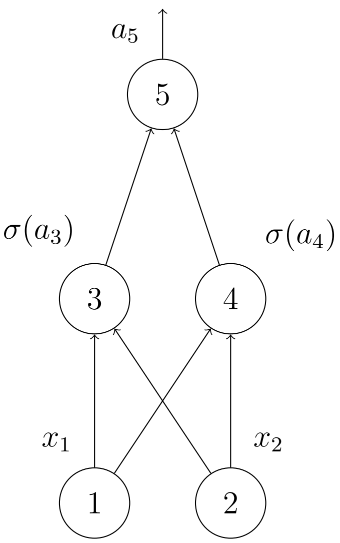
   (a) What is the probability that a randomly drawn point from p falls *outside* the interval $(z_\epsilon, \theta]$?

   (b) What is the probability that *all* N of the random examples in D fall outside the interval $(z_\epsilon, \theta]$?

   (c) What is the probability (over the randomly drawn training set) that A produces an $h_{\hat{\theta}}$ with error rate at least ϵ?

   (d) What is the smallest N (training set size) making that the probability (with respect to the random training examples) that $h_{\hat{\theta}}$ has error greater than ϵ is at most δ? (i.e. the probability that $h_{\hat{\theta}}$ has error less than ϵ is at least $1 - \delta$, express N as a function of ϵ and δ).

2. (3 pts) Consider the following artificial neural network.



Let the $\sigma()$ function be the logistic sigmoid, $\sigma(a) = 1/(1 + e^{-a})$. Assume that the the nodes do *not* have bias terms, and the initial weights are all 0's, the error on the output is the squared error, $\frac{1}{2}(a_5 - t)^2$, so $z_5 = a_5$ and there is no sigma-function at the output node. Under these assumptions, perform one step of backpropagation with step size $\eta = 0.1$ on the training example $x_1 = 1$, $x_2 = 2$, $y = 1$. Show the a_i, z_i and δ_i values for each non-input node, and the new weights after the backprop update. See the backpropagation handout for the procedure to use.