Foundations of Machine Learning
Department of Computer Science, NYU
Homework assignment 1
Due: February 8, 2005

Problem 2: Bernstein's Inequality

The objective of this problem is to prove Bennett's and Berstein's inequalities.

(1) Show that for any $t > 0$, and any random variable X with $\mathrm{E}[X] = 0$, $\mathrm{E}[X^2] = \sigma^2$, and $X \leq c$,

$$\mathrm{E}[e^{tX}] \leq e^{f(\sigma^2/c^2)},$$

where

$$f(x) = \log(\frac{1}{1+x}e^{-ctx} + \frac{x}{1+x}e^{ct}).$$

(2) Show that $f''(x) \leq 0$ for $x \geq 0$.

(3) Using Chernoff's bounding technique, show that

$$\Pr[\frac{1}{m}\sum_{i=1}^{m} X_i \geq \epsilon] \leq e^{-tm\epsilon + \sum_{i=1}^{m} f(\sigma_{X_i}^2/c^2)},$$

where ($\sigma_{X_i}^2$ is the variance of X_i.

(4) Show that $f(x) \leq f(0) + xf'(0) = (e^{ct} - 1 - ct)x$.

(5) Using the bound derived in (4), find the optimal value of t.

(6) *Bennett's inequality.* Let $X_1, \ldots, X_m$ be independent real-valued random variables with zero mean such that for $i = 1, \ldots, m$, $X_i \leq c$. Let $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} \sigma_{X_i}^2$. Show that

$$\Pr[\frac{1}{m}\sum_{i=1}^{m} X_i > \epsilon] \leq \exp(-\frac{m\sigma^2}{c^2}\theta(\frac{\epsilon c}{\sigma^2})),$$

where $\theta(x) = (1+x)\log(1+x) - x$.

(7) *Berstein's inequality.* Show that under the same conditions as Bennett's inequality

$$\Pr[\frac{1}{m}\sum_{i=1}^{m} X_i > \epsilon] \leq \exp(-\frac{m\epsilon^2}{2\sigma^2 + 2c\epsilon/3}).$$

Hint: show that for all $x \geq 0$, $\theta(x) \geq h(x) = \frac{3}{2}\frac{x^2}{x+3}$.

(8) Write Hoeffding's inequality assuming the same conditions. For what values of σ is Bernstein's inequality better than Hoeffding's inequality?

Problem 3: Two-Oracle Variant of PAC model

Assume that positive and negative examples are now drawn from two separate distributions D_+ and D_-. For an accuracy $(1-\epsilon)$, the learning algorithm must find a hypothesis h such that:

$$\Pr_{x\sim D_+}[h(x) = 0] \leq \epsilon \text{ and } \Pr_{x\sim D_-}[h(x) = 1] \leq \epsilon$$

Thus, the hypothesis must have a small error on both distributions.

Let C be any concept class and H be any hypothesis space. Let h_0 and h_1 represent the identically 0 and identically 1 functions, respectively. Prove that C is efficiently PAC-learnable using H in the standard (one-oracle) PAC model if and only if it is efficiently PAC-learnable using $H \cup \{h_0, h_1\}$ in this two-oracle PAC model.