Foundations of Machine Learning
Department of Computer Science, NYU
Homework assignment 1 – Solution

1. Bernstein's Inequality [40 points]

   (1) [20 bonus points]

   (2) [10 points] Just a series of calculations of the derivatives starting from:

   $$\forall x \geq 0, f'(x) = \frac{(-cte^{-ctx} + e^{ct})(1+x) - e^{-ctx} - xe^{ct}}{(1+x)^2} \frac{1+x}{e^{-ctx} + xe^{ct}}.$$

   This can be simplified into:

   $$\forall x \geq 0, f'(x) = \frac{e^{ct(x+1)} - (ctx + ct + 1)}{xe^{ct(x+1)} + 1}.$$

   The calculation of the second derivative leads to:

   $$\forall x \geq 0, f''(x) = -\frac{e^{2ct(x+1)} + c^2t^2x^2 + (c^2t^2 + 3ct)x + ct}{(xe^{ct(x+1)} + 1)^2} \leq 0.$$

   (3) [5 points] As already done in class in other instances, using Markov's inequality, for any $t > 0$,

   $$\Pr[X \geq m\epsilon] = \Pr[e^{tX} \geq e^{tm\epsilon}] \leq e^{-tm\epsilon}\mathrm{E}[e^{tX}].$$

   Using the inequality of (1) with $X = \sum_{i=1}^m X_i$ leads directly the desired inequality.

   (4) [10 points] By the Taylor series expansion with remainder, there exists $\theta \in [0, x]$ such that:

   $$f(x) = f(0) + xf'(x) + \frac{x^2}{2} f''(\theta)$$

   By (2), $f''(\theta) \leq 0$, thus $f(x) \leq f(0) + xf'(x)$.

   (5) [5 points] Plugging in the expression obtained in (3) in the inequality of (4) gives:

   $$\Pr[\frac{1}{m}\sum_{i=1}^m X_i \geq \epsilon] = \exp[-m\Phi(t)]$$

with $\Phi(t) = t\epsilon - (e^{ct} - 1 - ct)\frac{\sigma^2}{c^2}$. It is easy to see that:

$$\Phi'(t) \geq 0 \Leftrightarrow t \leq t_0 = \frac{1}{c}\log(1 + \frac{\epsilon c}{\sigma^2}).$$

Thus, t_0 is the optimal value.

(6) Replacing t by t_0 leads directly to Bennett's inequality.

(7) [5 points] It is sufficient to observe that: $\theta(0) = h(0) = 0$, $\theta'(0) = h'(0) = 0$, and $\forall x, \theta''(x) \geq h''(x)$.

$$\theta''(x) = \frac{1}{1+x} \text{ and } h''(x) = \frac{27}{(x+3)^3}$$

(8) [5 points] When $\mathrm{E}[X_i] = 0$ and $|X| \leq c$, Hoeffding's inequality (see also lemma proved in class) gives:

$$\Pr[\frac{1}{m}\sum_{i=1}^{m} X_i > \epsilon] \leq e^{-\frac{m\epsilon^2}{2c^2}}.$$

For smaller values of the variance, $\sigma^2 \ll c^2$, Bernstein's inequality is tighter.

2. Two-Oracle Variant of PAC model [60 points]

- [20 points] Assume that C is efficiently PAC-learnable using H in the standard PAC model using algorithm L. Consider the distribution $D = \frac{1}{2}(D_- + D_+)$. Let $h \in H$ be the hypothesis output by L. Choose δ such that:

$$\Pr[error_D(h) \leq \epsilon/2] \geq 1 - \delta.$$

From

$$\begin{aligned} error_D(h) &= \Pr_{x \sim D}[h(x) \neq c(x)] \\ &= \frac{1}{2}(\Pr_{x \sim D_-}[h(x) \neq c(x)] + \Pr_{x \sim D_+}[h(x) \neq c(x)]) \\ &= \frac{1}{2}(error_{D_-}(h) + error_{D_+}(h)), \end{aligned}$$

it follows that:

$$\Pr[error_{D_-}(h) \leq \epsilon] \geq 1 - \delta \quad \text{and} \quad \Pr[error_{D_+}(h) \leq \epsilon] \geq 1 - \delta.$$

This implies two-oracle PAC-learning with the same computational complexity.

- [40 points] Assume now that C is efficiently PAC-learnable in the two-oracle PAC model. Thus, there exists a learning algorithm L such that for $c \in C$, $\epsilon > 0$, and $\delta > 0$, there exist m_- and m_+ polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, such that if we draw m_- negative examples or more and m_+ positive examples or more, with confidence $1 - \delta$, the hypothesis h output by L verifies:

$$\Pr[error_{D_-}(h)] \leq \epsilon \quad \text{and} \quad \Pr[error_{D_+}(h)] \leq \epsilon.$$

Now, let D be a probability distribution over negative and positive examples. If we could draw m examples according to D such that $m \geq \max\{m_-, m_+\}$, m polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, then two-oracle PAC-learning would imply standard PAC-learning:

$\Pr[error_D(h)] \leq \Pr[error_D(h)|c(x) = 0] \Pr[c(x) = 0] +$
$\Pr[error_D(h)|c(x) = 1] \Pr[c(x) = 1] \leq$
$\epsilon(\Pr[c(x) = 0] + \Pr[c(x) = 1]) = \epsilon.$

If D is not too biased, that is if the probability of drawing a positive example, or that of drawing a negative example is more than ϵ, it is not hard to show, using Chernoff bounds or just Chebyshev's inequality, that drawing a polynomial number of examples in $1/\epsilon$ and $1/\delta$ suffices to guarantee that $m \geq \max\{m_-, m_+\}$ with high confidence.
Otherwise, D is biased towards negative (or positive examples), in which case returning $h = h_0$ (respectively $h = h_1$) guarantees that $\Pr[error_D(h)] \leq \epsilon$.
To show the claim about the not-too-biased case, let S_m denote the number of positive examples obtained when drawing m examples when the probability of a positive example is ϵ. By Chernoff bounds,

$$\Pr[S_m \leq (1 - \alpha)m\epsilon] \leq e^{-m\epsilon\alpha^2/2}.$$

We want to ensure that at least m_+ examples are found. With $\alpha = \frac{1}{2}$ and $m = \frac{2m_+}{\epsilon}$,

$$\Pr[S_m > m_+] \leq e^{-m_+/4}.$$

Setting the bound to be less than or equal to $\delta/2$, leads to the following condition on m:

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{8}{\epsilon} \log \frac{2}{\delta}\}$$

A similar analysis can be done in the case of negative examples. Thus, when D is not too biased, with confidence $1-\delta$, we will find at least m_- negative and m_+ positive examples if we draw m examples, with

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{2m_-}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}.$$