Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 3
April 14, 2011
Due: May 02, 2011

A. Kernels

1. Show that for all $\lambda > 0$, the kernel K defined on $\mathbb{R} \times \mathbb{R}$ by

$$K(x, x') = \exp\big(-\lambda[\sin(x' - x)]^2\big), \tag{1}$$

for all $x, x' \in \mathbb{R}$ is PDS (*hint*: you could seek to rewrite $[\sin(x' - x)]^2$ as the square of the norm of the difference of two vectors).

2. Let $\Phi\colon X \to H$ be a feature mapping such that the dimension N of H is very high and let $K\colon X \times X \to \mathbb{R}$ be a PDS kernel defined by

$$K(x, x') = \mathop{\mathrm{E}}_{i \sim D}\big[[\Phi(x)]_i [\Phi(x')]_i\big], \tag{2}$$

where $[\Phi(x)]_i$ is the ith componnent of $\Phi(x)$ and similarly for $\Phi(x')$ and where D is a distribution over the indices i. We shall assume that $|[\Phi(x)]_i| \leq R$ for all $x \in X$ and $i \in [1, N]$.

Suppose that to compute $K(x, x')$ no other method is available than computing the inner product (2), which would require $O(N)$ time. One idea in that case is instead to compute an approximation of that kernel based on random selection of a subset I of the N components of $\Phi(x)$ and $\Phi(x')$ according to D, that is:

$$K'(x, x') = \frac{1}{n} \sum_{i \in I} [\Phi(x)]_i [\Phi(x')]_i, \tag{3}$$

where $|I| = n$.

   (a) Fix x and x' in X. Prove that

$$\Pr_{I \sim D^n}[|K(x, x') - K'(x, x')| > \epsilon] \leq 2e^{\frac{-n\epsilon^2}{2R^4}}. \tag{4}$$

   (*hint*: use McDiarmid's inequality).

   (b) Let $\mathbf{K}$ and $\mathbf{K}'$ be the kernel matrices associated to K and K'. Show that for any $\epsilon, \delta > 0$, for $n > \frac{2R^4}{\epsilon^2} \log \frac{m(m+1)}{\delta}$, with probability at least $1 - \delta$, $|\mathbf{K}'_{ij} - \mathbf{K}_{ij}| \leq \epsilon$ for all $i, j \in [1, m]$.

B. Boosting

1. Let corr$(\mathbf{x}, \mathbf{x}')$ denote the inner product (or unnormalized correlation) of two vectors $\mathbf{x}$ and $\mathbf{x}'$. Prove that the distribution vector $(D_{t+1}(1), \ldots, D_{t+1}(m))$ defined by AdaBoost and the vector of components $y_i h_t(x_i)$ are uncorrelated.

2. Fix $\epsilon \in (0, 1/2)$. Let the training sample be defined by m points in the plane with $\frac{m}{4}$ negative points all at coordinate $(1, 1)$, another set of $\frac{m}{4}$ negative points all at coordinate $(-1, -1)$, $\frac{m(1-\epsilon)}{4}$ positive points all at coordinate $(1, -1)$, and $\frac{m(1+\epsilon)}{4}$ positive points all at coordinate $(-1, +1)$. Describe the behavior of AdaBoost when run on this sample using boosting stumps, in particular, give the solution the algorithm returns after T rounds.