Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 3 - solution
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).

For all $x, x' \in \mathbb{R}$,

$$\begin{aligned}[\sin(x' - x)]^2 &= 1 - [\cos(x' - x)]^2\\ &= 1 - [\cos x' \cos x + \sin x' \sin x]^2\\ &= 1 - (\mathbf{u}(x') \cdot \mathbf{u}(x))^2,\end{aligned}$$

where $\mathbf{u}(x) = (\cos x, \sin x)^\top$ for all $x \in X$. Observe that $\|\mathbf{u}(x)\| = 1$ for all $x \in X$. Thus, $[\sin(x' - x)]^2 = \frac{1}{2}\|\mathbf{u}(x') - \mathbf{u}(x)\|^2$ and $K(x, x') = e^{-\frac{\lambda}{2}\|\mathbf{u}(x') - \mathbf{u}(x)\|^2}$. Since the Gaussian kernel is known to be PDS, K is also PDS (the fact that Gaussian kernels are PDS can be shown easily by observing that they are the normalized kernels associated to the kernel obtained by composing $\exp$ with standard inner products).

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

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

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).

By definition of K', for all $x, x' \in X$, $\mathrm{E}_D[K'(x, x')] = K(x, x')$. Replacing one indice $i \in I$ by i' affects $K'(x, x')$ by at most $\frac{1}{n}(|[\Phi(x)]_i[\Phi(x')]_i| + |[\Phi(x)]_{i'}[\Phi(x')]_{i'}|) \leq \frac{2R^2}{n}$. The result thus follows directly 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 $i, j \in [1, m]$.

Since $\mathbf{K}$ and $\mathbf{K}'$ are symmetric, it suffices to prove the statement for the entries of these matrices that are above the diagonal. By the union bound and the previous question, the following holds:

$$\Pr_{I \sim D^n}[\exists i, j \in [1, m] \colon |\mathbf{K}'_{ij} - \mathbf{K}_{ij}| > \epsilon] \leq 2\frac{m(m+1)}{2} e^{\frac{-n\epsilon^2}{2R^4}} = m(m+1)e^{\frac{-n\epsilon^2}{2R^4}}.$$

Setting $\delta > 0$ to match the upper bound leads directly to the inequality claimed.

B. Boosting

1. Let $\mathrm{corr}(\mathbf{x}, \mathbf{x}')$ 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.

By definition, the unnormalized correlation is given by

$$\sum_{i=1}^{m} D_{t+1}(i) y_i h_t(x_i) = \sum_{i=1}^{m} \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)} y_i h_t(x_i)}{Z_t} = \frac{1}{Z_t} \frac{dZ_t}{d\alpha_t}, \tag{5}$$

since $Z_t = \sum_{i=1}^m D_t(i) e^{-\alpha_t y_i h_t(x_i)}$. Recall from lecture slides that α_t minimizes Z_t, thus $\frac{dZ_t}{d\alpha_t} = 0$. This shows that the distribution at round $t+1$ and the vector of margins at round t 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.

   It is not hard to see that the base hypotheses in this problem can be defined to be threshold functions based on the first or second axis, or constant functions (horizontal or vertical thresholds outside the convex hull of all the points).

   The hypotheses selected by AdaBoost are therefore chosen from this set. It can be shown that the hypotheses selected in two consecutive rounds of AdaBoost are distinct. Furthermore, h_t and $-h_t$ cannot be selected in consecutive rounds since misclassified and correctly classified points by h_t are assigned the same distribution mass (see lecture slides). Thus, at each round a distinct hypothesis is chosen. The points at coordinate $(-1, -1)$ are misclassified by all these base hypothesis.

   The algorithm should be stopped when the best ϵ_t found is $1/2$. It can be shown then that the error of the final classifier returned on the training set is $\frac{1}{4}(1 - \epsilon)$ since it misclassifies exactly the points at at coordinate $(+1, -1)$.