UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281B / STAT 241B
ADVANCED TOPICS IN STATISTICAL LEARNING THEORY

Problem Set 2

Spring 2009

Issued: Monday, February 9, 2009 **Due:** Monday, February 23, 2009

Problem 2.1
True or false: either provide a proof (when true) or an explicit counterexample (when false).

(a) If $\mathbb{K}_1$ and $\mathbb{K}_2$ are both positive semidefinite (PSD) kernel functions on $\mathcal{X} \times \mathcal{X}$, then $\lambda_1 \mathbb{K}_1 + \lambda_2 \mathbb{K}_2$ is a PSD kernel function for all $\lambda_i \geq 0$.

(b) Any symmetric function $\mathbb{K}$ is that is elementwise non-negative (i.e., $\mathbb{K}(x, y) \geq 0$ for all x, y) is a PSD kernel function.

(c) If $\mathbb{K}_1$ and $\mathbb{K}_2$ are both PSD kernel functions on $\mathcal{X} \times \mathcal{X}$, then $\mathbb{K}(x, y) := \mathbb{K}_1(x, y)\mathbb{K}_2(x, y)$ is also a PSD kernel function.

(d) Given a probability space with events $\mathcal{E}$ and probability law $\mathbb{P}$, the function $\mathbb{K} : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}$ defined by $\mathbb{K}(A, B) := \mathbb{P}(A, B) - \mathbb{P}(A)\mathbb{P}(B)$ is a PSD kernel function.

(e) Given a finite set $\mathcal{E}$, let $\mathcal{P}(\mathcal{E})$ denote the set of all subsets of $\mathcal{E}$. If $\mathbb{K} : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}$ is a PSD kernel function, then

$$\bar{\mathbb{K}}(A, B) \quad := \quad \sum_{x \in A, y \in B} \mathbb{K}(x, y)$$

is a PSD kernel function on $\mathcal{P}(\mathcal{E}) \times \mathcal{P}(\mathcal{E})$.

Problem 2.2
On the course website, you will find the data set `regression.dat` in ASCII format, which defines a regression problem in $\mathbb{R}^{10}$. (The first 10 columns correspond to $(x_1, \ldots, x_{10})$ and the final column corresponds to $y \in \mathbb{R}$.)

(a) Fit a linear regression to these data and report the sum of squared errors on the test set `regression.test`.

(b) Use ordinary PCA and reduce the dimensionality of the covariate space to two dimensions. Fit a linear regression and report the sum of squared errors on the test set `regression.test`.

(c) Use kernel PCA with a Gaussian kernel $K(x, y) = \exp(\frac{-\|x-y\|^2}{2\sigma^2})$, and reduce the dimensionality of the covariate space to two dimensions. (Propose and implement a method for choosing the bandwidth parameter σ). Fit a linear regression and report the sum of squared errors on the test set `regression.test`.

Problem 2.3
For each of the following kernels, compute the eigenfunctions and eigenvalues of the operator $T_{\mathbb{K}} : L^2(\mathcal{E}) \to L^2(\mathcal{E})$ defined by

$$T_{\mathbb{K}}(f)(x) \quad = \quad \int_{\mathcal{E}} \mathbb{K}(x,y) f(y) dy.$$

(a) For $\mathcal{E} = [0, 2\pi]$, the kernel $\mathbb{K}(x,y) = \sum_{\ell=0}^{\infty} w_\ell \cos(\ell(x-y))$ for some sequence of weights $w_\ell \geq 0$ such that $\sum_{\ell=0}^{\infty} w_\ell < \infty$.

(b) For $\mathcal{E} = [0,1]$, the polynomial kernel $K(x,y) = (1+xy)^2$.

Problem 2.4
Consider a RKHS with feature map Φ and kernel $\mathbb{K}$, such that $\mathbb{K}(x,y) = \langle \Phi(x),\, \Phi(y)\rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{E}$. Given a data set $\{x^{(1)}, \ldots, x^{(n)}\}$, consider some element f in the linear span of $\{\Phi(x^{(i)}), i = 1, 2, \ldots, n\}$—that is, $f = \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)})$ for some fixed coefficients $\alpha \in \mathbb{R}^n$. The projection of a new element $\Phi(x)$ onto f is given by

$$\frac{\langle f,\, \Phi(x)\rangle_{\mathcal{H}}}{\|f\|_{\mathcal{H}}^2} f.$$

Show how to compute the sample variance of this projection, for a fixed $\Phi(x)$, using only the kernel $\mathbb{K}$.

Problem 2.5
Given a data set $\{x^{(1)}, \ldots, x^{(n)}\} \subseteq \mathbb{R}^d$, a novelty detection algorithm can be constructed by finding the smallest sphere that contains the data points. (When a new x is observed, it is flagged as "novel" if it lies outside this sphere.) Of course, this idea can also be implemented in a feature space, using some feature map Φ associated with a RKHS (i.e., $\mathbb{K}(x,y) = \langle \Phi(x),\, \Phi(y)\rangle_{\mathcal{H}}$ for all $x, y \in \mathbb{R}^d$).

(a) Give a precise formulation of the optimization problem to be solved in order to learn a novelty detector. Using Lagrangian methods, compute the dual, and show how solution requires only computing the kernel matrix K with entries $K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)})$.

(b) Extend your algorithm to allow some fraction $\nu > 0$ of the data to allow outside the sphere in feature space. (*Hint:* Use slack variables, as in the extension of a hard margin SVM to a soft margin SVM.)

Problem 2.6
Concentration bounds play an important role in the analysis of statistical estimators; in this problem, we explore some elementary aspects of concentration.

(a) Prove that if Z is a non-negative random variable with expectation $\mathbb{E}[Z]$, then for all $t > 0$, we have $\mathbb{P}[Z \geq t] \leq \mathbb{E}[Z]/t$.

(b) A zero-mean random variable is said to be sub-Gaussian with parameter $\sigma > 0$ if $\mathbb{E}[\exp(sX)] \leq \exp(\frac{\sigma^2 t^2}{2})$ for all $s \in \mathbb{R}$. Show that $X \sim N(0, \sigma^2)$ is sub-Gaussian.

(c) Suppose that X is Bernoulli with $\mathbb{P}[X = +1] = \mathbb{P}[X = -1] = 1/2$. Show that X is sub-Gaussian. (Can you generalize your argument to any bounded random variable?)

(d) Show that any sub-Gaussian random variable X satisfies the two-sided tail bound

$$\mathbb{P}[|X| > t] \;\leq\; 2\exp\big(\frac{-t^2}{2\sigma^2}\big) \qquad \text{for all } t \in \mathbb{R}.$$

(e) Let $X_1, \ldots, X_n$ be n i.i.d. samples of a sub-Gaussian variable with parameter σ. Show that for any $\delta > 0$, we have

$$\mathbb{P}\big[\max_{i=1,\ldots,n} X_i > \sqrt{(2+\delta)\sigma^2 \log n}\big] \;\rightarrow\; 0 \quad \text{as } n \rightarrow +\infty.$$