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

EECS 281B / STAT 241B
Advanced Topics in Statistical Learning Theory

Problem Set 1
Spring 2009

Issued: Monday, January 26, 2009 **Due:** Monday, February 9, 2009

Notation: Throughout this problem set, we use the definition $\eta(x) = \mathbb{P}[Y = 1 \mid X = x]$.

Problem 1.1
The *Bayes risk* of a binary classification problem under $0-1$ loss is given by

$$R^* \;=\; \inf_{g:\mathcal{X}\to\{-1,+1\}} \mathbb{P}[g(X) \neq Y].$$

Prove the following equivalent representations of the Bayes risk:

(a) $R^* = \mathbb{E}[\min\{\eta(X), 1-\eta(X)\}]$.

(b) $R^* = \frac{1}{2} - \frac{1}{2}\mathbb{E}\big|2\eta(X) - 1\big|$.

(c) Suppose that for $y \in \{-1,+1\}$, the random variables $(X \mid Y = y)$ have densities, say f_+ and f_- for $y = +1$ and $y = -1$ respectively, and that $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1] = 1/2$. Show that

$$R^* \;=\; \frac{1}{2} - \frac{1}{4}\int \big|f_+(x) - f_-(x)\big| dx.$$

Problem 1.2
Following up on the example from class, consider the problem of classifying pass/fail performance ($Y = +1$ means pass) in a class based on (X_1, X_2, X_3), where $X_1 = \#$ hours sleeping, $X_2 = \#$ hours playing video games and $X_3 =$ laziness measure. We assume that $\eta(x) = \mathbb{I}[x_1 + x_2 + x_3 \leq 7]$.

(a) Find the Bayes classifier and Bayes risk (under $0-1$ loss) for classification based on (X_1, X_2) when the triple (X_1, X_2, X_3) are uniformly distributed over $[0,4]^3$.

(b) Let (X_1, X_2, X_3) be independent. Show that by changing only the distribution of X_3, the Bayes error for classification based on (X_1, X_2) can be made arbitrarily close to $1/2$.

(c) Let X_1 and X_2 be independent, each with Exp(1) distributions. Find a joint distribution of (X_1, X_2, X_3) such that the Bayes classifier g^* based on (x_1, x_2) is given by

$$g^*(x_1, x_2) \;=\; \begin{cases} +1 & \text{if } x_1^2 + x_2^2 < 10 \\ -1 & \text{otherwise.} \end{cases}$$

Problem 1.3

(a) Implement a linear perceptron algorithm, using the initialization $\theta^0 = 0$. Apply it to the ASCII format data files `X.dat` and `Y.dat` from the webpage. Compare the number of iterations required to the theoretical bound derived in class. (You can use the fixed point θ^* that you obtain to estimate the margin δ.)

(b) Implement and apply the linear SVM to the same data set. (In MATLAB, you can use the function `quadprog.m` to solve a quadratic program.)

Problem 1.4

The Kullback-Leibler divergence between two random vectors X_1 and X_2 with density functions f_1 and f_2 over a common support $S \subseteq \mathbb{R}^d$ is given by $D(f_1 \| f_2) = \int_S f_1(x) \log \frac{f_1(x)}{f_2(x)} dx$.

(a) Compute the Kullback-Leibler divergence for $X_i \sim N(\mu_i, \Sigma_i)$ for $i = 1, 2$.

(b) Suppose that for $y \in \{-1, +1\}$, $(X \mid Y = y) \sim N(\mu_y, \Sigma)$, and that $\mathbb{P}[Y = +1] = p$. Prove that

$$\mathbb{E}[\sqrt{\eta(X)(1-\eta(X))}] \quad = \quad \sqrt{p\,(1-p)}\exp(-\frac{1}{4}D(\mu_{+1}\|\mu_{-1}))$$

where $D(\mu_{+1}\|\mu_{-1})$ is the Kullback-Leibler divergence between $N(\mu_{+1}, \Sigma)$ and $N(\mu_{-1}, \Sigma)$.

Problem 1.5

In lecture, we considered the linear maximum margin classifier, working under the assumption that the data set was linearly separable. In practice, this condition need not hold, so that it is more natural to consider a MM classifier with non-negative slack variables $\{\upsilon_i, i = 1, \ldots, n\}$ that permit some violations. Consider the classifier based on solving the modified quadratic program:

$$\begin{array}{ll} \min_{\theta \in \mathbb{R}^d, \upsilon \in \mathbb{R}^n_+} & \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^n \upsilon_i \\ \text{such that} & y^{(i)}\langle \theta,\, x^{(i)}\rangle \geq 1 - \upsilon_i \qquad \forall i = 1, \ldots, n \end{array}$$

Here $C > 0$ is a constant, to be chosen by the user.

(a) Derive the dual function $q(\alpha) = \inf_{\theta \in \mathbb{R}^d, \upsilon \in \mathbb{R}^n_+} L(\theta, \upsilon; \alpha)$ associated with this quadratic program.

(b) Suppose that we add a constant term to our classifier: i.e., we consider linear functions of the form $\langle (\theta, b),\, (x, 1)\rangle = \langle \theta,\, x\rangle + b$ where $b \in \mathbb{R}$ is an offset to be optimized. Derive the dual function associated with this modified classifier.

Problem 1.6

Given a non-empty set $\mathcal{S}$, a function $\mathbb{K} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is a positive semidefinite kernel function if for any natural number n and vectors $\{x^{(i)}, i = 1, \ldots, n\}$ from $\mathcal{S}$, the matrix $K \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)})$ is symmetric, and positive semidefinite.

(a) For any kernel function $\mathbb{K}$ and vectors $x^{(1)}, x^{(2)} \in \mathcal{S}$, show that

$$|\mathbb{K}(x^{(1)}, x^{(2)})|^2 \quad \leq \quad \mathbb{K}(x^{(1)}, x^{(1)})\, \mathbb{K}(x^{(2)}, x^{(2)}).$$

(b) Suppose that there exists a function $\Phi : \mathcal{S} \to \mathbb{R}^D$ such that for all $x, y \in \mathcal{S}$, we have $\mathbb{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Show that $\mathbb{K}$ is a positive semidefinite kernel function.

(c) Show that for $\mathcal{S} = \mathbb{R}^d$, the function $\mathbb{K}(x, y) = (\langle x, y \rangle)^3$ is a kernel function.

(d) Given a set Ω, let $\mathcal{S} = \mathcal{P}(\Omega)$—that is, the set of all subsets of Ω. Define a function on $\mathcal{S} \times \mathcal{S}$ via $\mathbb{K}(A, B) = 2^{|A \cap B|}$. Show that $\mathbb{K}$ is a positive semidefinite kernel function.