

Midterm solutions

Wednesday, April 8, 2009

Problem:

[16 pts, 4 pts each] Consider the indicator kernel function

$$\mathbb{K}(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise.} \end{cases}$$

Consider a set of samples $\{x^{(i)}, y^{(i)}\}_{i=1}^n$, where each pair $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \{-1, +1\}$, and assume that all the feature vectors $x^{(i)}$ are distinct. Now suppose that we estimate a classifier f by optimizing over the RKHS defined by \mathbb{K} , using the hard-margin support vector machine approach.

- (a) Explain why the optimal solution can be written in the form

$$\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i y^{(i)} \mathbb{K}(\cdot, x^{(i)})$$

for a vector $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n)$ of data-dependent weights.

- (b) Specify the dual program that specifies the optimal weight vector $\hat{\alpha}$, and compute the optimal weight vector $\hat{\alpha}$ for the indicator kernel. (*Hint:* It can be solved explicitly for the special case of the indicator kernel.) Specify the number of support vectors in your solution.
- (c) Now suppose that we use the optimal solution \hat{f} to make predictions via the function $g: \mathbb{R}^d \rightarrow \{-1, +1\}$ given by

$$g(x) = \text{sign}(\hat{f}(x)) \in \{-1, +1\}.$$

(For concreteness, say that $\text{sign}(0) = +1$). What is your classification on any one of the training samples $x^{(i)}$?

- (d) Suppose that we draw a random sample (X, Y) from a distribution with $\mathbb{P}[Y = +1] = q$, for some $q \in [0, 1]$, and with X a continuous random vector with a density function. Can you compute the classification error $\mathbb{P}[Y \neq g(X)]$ for your classifier g applied to this random sample?

Solution:

- (a) By definition, the hard-margin SVM solves the optimization problem $\min_{f \in \mathcal{H}} \frac{1}{2} \|f\|_{\mathcal{H}}^2$ subject to $y^{(i)} f(x^{(i)}) \geq 1$ for all $i = 1, \dots, n$. Equivalently, it minimizes the functional

$$J(f) = \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \sum_{i=1}^n G(y^{(i)} f(x^{(i)}) - 1),$$

where $G(t) = 0$ if $t \geq 0$ and $G(t) = +\infty$ otherwise. By the representer theorem, the solution takes the given form.

- (b) From class and homework, we know that the hard-margin SVM over a RKHS, when dualized, is equivalent to solving the quadratic program

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \alpha^T (K \odot Y) \alpha$$

where $K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)})$ and $Y_{ij} = y^{(i)} y^{(j)}$. Given the special form of the indicator kernel, in fact we have $K \odot Y = I_{n \times n}$, so that the dual hard-margin SVM program becomes $\min_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \|\alpha\|_2^2$. Taking derivatives yields $\hat{\alpha} = \mathbf{1}$. Consequently, every data point is a support vector for this problem.

Another way to obtain the dual is to substitute the form of optimal solution $\hat{f}(\cdot)$ back to the functional $J(f)$ from part (a), and minimize it with respect to α . Doing so yields the dual problem $\min_{\alpha \in \mathbb{R}^d} \frac{1}{2} \alpha^T (K \odot Y) \alpha$ subject to $\alpha_i \geq 1$ for all $i = 1, \dots, n$. Directly solving this problem give us $\hat{\alpha} = \mathbf{1}$, as above.

- (c) From part (a) and class, we know that

$$\begin{aligned} \hat{f}(x) &= \sum_{i=1}^n \hat{\alpha}_i y^{(i)} \mathbb{K}(x, x^{(i)}) \\ &= \begin{cases} y^{(i)} & \text{if } x = x^{(i)} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- (d) From part (c), the classifier will output $\hat{f}(x) = 0$ for any sample that is not one of the training samples. Since X is continuous, with probability one, a randomly drawn sample is distinct from the training set so that $g(X) = \text{sign}(\hat{f}(X)) = +1$ with probability 1. Thus, the test error will be equal to $\mathbb{P}[Y \neq +1] = 1 - q$.

Problem:

[20 pts, 4 pts each] True or false: either provide a proof (when true) or an explicit counterexample (when false). (**Note:** You will receive zero points for only stating true or false without justification.)

- (a) Let X and Y be independent zero-mean sub-Gaussian random variables. Then $aX + bY$ is also sub-Gaussian for any real constants a and b .
- (b) The class of sets \mathcal{A} given by all convex polygons in \mathbb{R}^2 has finite VC dimension.
- (c) Let $\mathcal{X} \subset \mathbb{R}^d$ and let $\Phi : \mathcal{X} \rightarrow \mathcal{H}$, where \mathcal{H} is some Hilbert space. (I.e., for each point $x \in \mathcal{X}$, the quantity $\Phi(x)$ is a member of the Hilbert space \mathcal{H} .) Then the function

$$\mathbb{K}(x, y) = \frac{\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}}{\|\Phi(x)\|_{\mathcal{H}} \|\Phi(y)\|_{\mathcal{H}}}$$

is a positive semidefinite kernel function.

- (d) Consider a random vector (X_1, \dots, X_d) such that each X_i has mean zero and unit variance, $|X_i| \leq B/2$ for all $i = 1, \dots, d$, and $\mathbb{E}[X_i X_j] = \mu$ for $i \neq j$. Then the random variable $Z_d = \frac{1}{\binom{d}{2}} \sum_{i=1}^d \sum_{j=1, j \neq i}^d X_i X_j$ satisfies a bound of the form

$$\mathbb{P}[|Z_d - \mathbb{E}[Z_d]| \geq t] \leq c_1 \exp(-c_2 dt^2) \quad \text{for some constants } c_1 \text{ and } c_2.$$

- (e) Given two positive semidefinite kernel functions \mathbb{K}_1 and \mathbb{K}_2 , the function

$$\mathbb{K}(x, y) = \min\{\mathbb{K}_1(x, y), \mathbb{K}_2(x, y)\}$$

is also a positive semidefinite kernel function.

Solution:

- (a) TRUE. Since X and Y are sub-Gaussian, there are constants σ_X and σ_Y such that $\mathbb{E}[\exp(tX)] \leq \exp(t^2 \sigma_X^2 / 2)$ and $\mathbb{E}[\exp(tY)] \leq \exp(t^2 \sigma_Y^2 / 2)$ for all $t \in \mathbb{R}$. Therefore, using independence, we have

$$\begin{aligned} \mathbb{E}[\exp(t(aX + bY))] &= \mathbb{E}[\exp(taX)] \mathbb{E}[\exp(tbY)] \\ &\leq \exp((a^2 \sigma_X^2 + b^2 \sigma_Y^2) t^2 / 2), \end{aligned}$$

which shows that $aX + bY$ is sub-Gaussian with parameter $\sqrt{a^2 \sigma_X^2 + b^2 \sigma_Y^2}$.

- (b) FALSE. For any positive integer n , suppose that we place the n points $\{x_1, x_2, \dots, x_n\}$ uniformly spaced on the unit circle. Then for each of the 2^n subsets of this data set, there is convex polygon that contains it, and excludes its complement. Thus, the shatter coefficient is 2^n for all n , implying that the VC dimension is infinite.
- (c) TRUE. We know that any function $\mathbb{K}(x, y)$ that is formed as a Gram matrix of the form $\langle f(x), f(y) \rangle$ for some function f is PSD. The given function has this representation with $f(x) = \Phi(x) / \|\Phi(x)\|_{\mathcal{H}}$, and taking inner products in the Hilbert space.

- (d) FALSE. Let X_1 be a discrete random variable with $\mathbb{P}_X(\sqrt{2}) = \mathbb{P}_X(-\sqrt{2}) = 1/4$ and $\mathbb{P}_X(0) = 1/2$, and set $X_i = X_1$ for all $i = 2, 3, \dots, d$. Then for all i and j , we have $\mathbb{E}[X_i] = 0$, $\text{var}(X_i) = 1$ and $\mathbb{E}[X_i X_j] = \mathbb{E}[X_1^2] = 1$, and hence $\mathbb{E}[Z_d] = \mathbb{E}[X_1^2] = 1$. However, with probability $1/2$, the random variable Z_d is equal to 0 , so that the concentration condition does not hold.

Only partial credit was given to those who stated TRUE, and tried to apply the bounded difference inequality, since it cannot be applied unless we assume that the X_i are independent (which would require $\mu = 0$, among other conditions).

- (e) FALSE. Consider the 2×2 matrices

$$K_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \text{and} \quad K_2 = \begin{bmatrix} 0.5 & 1 \\ 1 & 2 \end{bmatrix},$$

both of which are positive semidefinite. However, we have

$$K' = \min(K_1, K_2) = \begin{bmatrix} 0.5 & 1 \\ 1 & 1 \end{bmatrix},$$

which is not PSD.

Problem:

[20 pts] Given a fixed vector $\theta \in \mathbb{R}^{d+1}$, define the polynomial function $f : \mathbb{R} \rightarrow \mathbb{R}$ by

$$f_{\theta}(x) = \sum_{j=0}^d \theta_j x^j,$$

and consider the function class $\mathcal{F} := \{f_{\theta} \mid \theta \in [-1, +1]^{d+1}\}$.

(a) [6 pts] Derive an upper bound on the VC dimension of the class of sets

$$\mathcal{A}_{\mathcal{F}} := \{\{x \mid f_{\theta}(x) \geq 0\} \mid f_{\theta} \in \mathcal{F}\}.$$

(b) [4 pts] Derive an upper bound on the n th shatter coefficient $\mathcal{S}(n, \mathcal{A}_{\mathcal{F}})$ of this family.

(c) [5 pts] Given n samples $\{X^{(1)}, \dots, X^{(n)}\}$ where each $X^{(i)} \in \mathbb{R}$, consider the $n \times (d+1)$ matrix M formed as

$$M = \begin{bmatrix} 1 & X^{(1)} & (X^{(1)})^2 & \dots & (X^{(1)})^d \\ 1 & X^{(2)} & (X^{(2)})^2 & \dots & (X^{(2)})^d \\ \vdots & \vdots & & & \vdots \\ 1 & X^{(n)} & (X^{(n)})^2 & \dots & (X^{(n)})^d \end{bmatrix}$$

Assuming that M has rank n , compute the empirical Rademacher complexity $\widehat{\mathbb{R}}_n(\mathcal{G})$ of the class \mathcal{G} of functions $g : \mathbb{R} \rightarrow \{-1, +1\}$ given by

$$\mathcal{G} = \{\text{sign}(f(x)) \mid f_{\theta} \in \mathcal{F}\}.$$

(Consider the Rademacher complexity conditioned on the data $\{X^{(i)}\}_{i=1}^n$, so that $\widehat{\mathbb{R}}_n(\mathcal{G})$ is not random.)

(d) [5 pts] Now suppose that we specialize to quadratic functions (i.e., $d = 2$). Let \mathbb{P} be a Bernoulli distribution with $\mathbb{P}[V = 1] = 2/3$ and $\mathbb{P}[V = 0] = 1/3$, and let \mathbb{Q} be a discrete distribution with $\mathbb{Q}[U = 1/2] = 8/9$ and $\mathbb{Q}[U = 2] = 1/9$. Let $V^{(1)}, \dots, V^{(n)}$ be i.i.d. draws from \mathbb{P} , and let $U^{(1)}, \dots, U^{(n)}$ be i.i.d. draws from \mathbb{Q} . Consider the sequence of random variables

$$Z_n = \sup_{f_{\theta} \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f_{\theta}(V^{(i)}) - \frac{1}{n} \sum_{i=1}^n f_{\theta}(U^{(i)}) \right|.$$

Does the sequence $\{Z_n\}$ converge in probability to zero? Why or why not?

Solution:

(a) We observe that the space of polynomials of degree d in one variable is a vector space of dimension $d + 1$. Therefore, by a result proved in Lecture #14 on VC dimension of vector spaces, we know that $V_{\mathcal{F}} \leq d + 1$.

(b) From Sauer's lemma (stated in lecture), we know that $\mathcal{S}(n, \mathcal{F}) \leq (n+1)^{V_{\mathcal{F}}} \leq (n+1)^{d+1}$.

- (c) We observe that under the stated condition on the $n \times (d+1)$ matrix M , we can find a solution $\gamma \in \mathbb{R}^{d+1}$ to the linear system $M\gamma = \vec{\sigma}$, where $\vec{\sigma} \in \mathbb{R}^n$ is the $(n+1)$ vector of signs that define the Rademacher complexity. By rescaling as needed, we can find a vector $\theta^* \in [-1, +1]^{d+1}$ such that $\text{sign}(M\theta^*) = \sigma$. Consequently, for any choice of $\vec{\sigma}$, we have

$$\sup_{f_\theta \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \text{sign}(f_\theta(x)) \geq \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \text{sign}(f_{\theta^*}(x)) = 1.$$

Therefore, we have $\widehat{\mathbb{R}}_n(\mathcal{F}) = 1$.

- (d) We begin by observing under the given distributions \mathbb{P} and \mathbb{Q} , we have

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[U] &= \frac{1}{2} \frac{8}{9} + 2 \frac{1}{9} = 2/3, \quad \text{and} \\ \mathbb{E}_{\mathbb{Q}}[U^2] &= \frac{1}{4} \frac{8}{9} + 4 \frac{1}{9} = 2/3. \end{aligned}$$

Moreover, $\mathbb{E}_{\mathbb{P}}[X] = \mathbb{E}_{\mathbb{P}}[X^2] = 2/3$, so that by linearity of expectation, we have that $\mathbb{E}_{\mathbb{P}}[f_\theta(X)] = \mathbb{E}_{\mathbb{Q}}[f_\theta(U)]$ for all f_θ in the family. Therefore, we have

$$\begin{aligned} Z_n &= \sup_{f_\theta \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f_\theta(X^{(i)}) - \frac{1}{n} \sum_{i=1}^n f_\theta(U^{(i)}) \right| \\ &\leq \sup_{f_\theta \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f_\theta(X^{(i)}) - \mathbb{E}_{\mathbb{P}}[f_\theta(X)] \right| + \sup_{f_\theta \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f_\theta(U^{(i)}) - \mathbb{E}_{\mathbb{Q}}[f_\theta(U)] \right|, \end{aligned}$$

by applying triangle inequality. The problem is thus reduced to the uniform law of large numbers over \mathcal{F} , once for \mathbb{P} and once for \mathbb{Q} . Using part (b) and results from class, we know that

$$\mathbb{P} \left[\sup_{f_\theta \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f_\theta(X^{(i)}) - \mathbb{E}_{\mathbb{P}}[f_\theta(X)] \right| > t \right] \leq c_1(n+1)^3 \exp(-c_2 n t^2),$$

for some constants c_1 and c_2 , with a similar bound for the term involving U and \mathbb{Q} . Thus, we have shown that Z_n converges to zero in probability.