ECE 543: Statistical Learning Theory

Maxim Raginsky

March 25, 2021

Homework 3

Assigned March 25; due April 6, 2021

Note: natural logarithms are used throughout.

1. **Fast rates in binary classification.** In this problem, you will prove that the excess risk of ERM for binary classification can, in certain cases, be as low as $O(1/n)$, in contrast to the usual $O(1/\sqrt{n})$ behavior (here n is the size of the training set). For simplicity, we will only consider the case when the class $\mathcal{F}$ of candidate classifiers $f : \mathsf{X} \to \{0,1\}$ is a finite set.

   Thus, let $(X,Y) \in \mathsf{X} \times \{0,1\}$ be a random couple with distribution $P = P_{XY}$, and let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be n i.i.d. samples from P. Consider forming the usual empirical estimate of the loss $L(f) = \mathbf{P}(f(X) \neq Y)$ of every classifier $f \in \mathcal{F}$:

$$L_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f(X_i) \neq Y_i\}},$$

   so that the ERM solution is

$$\widehat{f}_n = \mathop{\arg\min}_{f \in \mathcal{F}} L_n(f) \equiv \mathop{\arg\min}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f(X_i) \neq Y_i\}}.$$

   (a) Prove that, for any $f \in \mathcal{F}$,

$$L(f) \le L_n(f) + \sqrt{\frac{2L(f)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}$$

   with probability at least $1-\delta$.

   Hint: You may need the following version of *Bernstein's inequality* — if $U_1,\ldots,U_n$ are n i.i.d. Bernoulli(p) random variables, then

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n U_i < p - \varepsilon\right) \le \exp\left(-\frac{n\varepsilon^2}{2p + 2\varepsilon/3}\right).$$

   (b) Use the result from part (a) to show that, for any $f \in \mathcal{F}$,

$$L(f) \le L_n(f) + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{n}$$

with probability at least $1-\delta$. Use this to prove that if the ERM solution classifies every training example correctly, i.e., if $L_n(\widehat{f}_n) = 0$, then

$$L(\widehat{f}_n) \leq \frac{4\log(|\mathcal{F}|/\delta)}{n}, \qquad \text{with probability at least } 1-\delta.$$

(In particular, this bound holds when the relationship between X and Y is deterministic, $Y = f(X)$, and the function f happens to lie in $\mathcal{F}$.)

Hint: You may need the fact that, for any three nonnegative numbers a, b, c, $a \leq b + c\sqrt{a}$ implies $a \leq b + c^2 + c\sqrt{b}$.

2. **VC dimension of combined classifiers using hard thresholding.** Let $\mathcal{G}$ denote the set of interval classifiers $g : \mathbb{R} \to \{1, -1\}$. Each $g \in \mathcal{G}$ has the form $g(x) = \text{sgn}((x-a)(b-x))$ for $a \leq b \in \mathbb{R}$, where $\text{sgn}(u) = \mathbf{1}_{\{u \geq 0\}} - \mathbf{1}_{\{\mathbf{u<0}\}}$.

   (a) What is the VC dimension, $V(\mathcal{G})$, and what is the resulting upper bound on the maximum Rademacher complexity for a sample of size n: $R_n(\mathcal{G}(x^n))$ for samples $\{x_1, \ldots, x_n\} \subset \mathbb{R}$, for $n \geq 1$ (obtained from the finite class lemma, Sauer-Shelah lemma, and $\binom{n}{\leq d} \leq (n+1)^d$) ?

   (b) Let $\mathcal{G}_1$ be the set of classifiers of the form $g(x) = \text{sgn}\left(\sum_{i=1}^{N} c_i g_i(x)\right)$, where $N \geq 1$, $g_i \in \mathcal{G}$ for $i \in [N]$, and $(c_1, \ldots, c_N)$ is a probability vector. Thus, g can be the result of comparing a convex combination of arbitrarily many simple interval classifiers to the threshold 0. In short, $\mathcal{G}_1 = \text{sgn}(\text{conv}(\mathcal{G}))$. Identify the VC dimension of $\mathcal{G}_1$ and the Rademacher average for n sample points, $R_n(\mathcal{G}_1(x^n))$ (with notation as in part (a)).

3. **Transformation of Mercer kernels.** Let $A \odot B$ denote Hadamard (i.e., elementwise) multiplication for two vectors or matrices of the same dimension. For example, $(A \odot B)_{ij} := A_{ij}B_{ij}$ for all i, j.

   (a) Suppose X and Y are two mean zero random vectors with values in $\mathbb{R}^d$. Denote their respective covariance matrices by $\Sigma_X = \mathbf{E}[XX^T]$ and $\Sigma_Y = \mathbf{E}[YY^T]$. Suppose X and Y are independent of each other. Express the covariance matrix of $X \odot Y$ in terms of Σ_X and Σ_Y.

   (b) Show that the product of two Mercer kernels for the same domain X is a Mercer kernel.

   Hint: A symmetric real matrix is positive semidefinite (PSD) if and only if it is the covariance matrix for some mean zero random vector.

4. **Half-space classifiers and support vector machines (SVM)**

   Consider the concept learning problem $(\mathsf{X} = \mathbb{R}^d, \mathsf{Y} = \{\pm 1\}, \mathcal{P}, \mathcal{G})$ with 0-1 loss, where $\mathcal{P}$ is a set of probability distributions P on $\mathsf{Z} = \mathsf{X} \times \{\pm 1\}$, and $\mathcal{G}$ consists of all half-space classifiers of the form $g_{w,b}(x) = \text{sgn}(\langle w, x\rangle + b)$, where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The generalization loss is defined by $L_P(w, b) = \mathbf{P}\{Y \neq \text{sgn}(\langle w, X\rangle + b)\}$.

   (a) Explain why this problem is PAC learnable. That is, describe a PAC learning algorithm and give a performance bound demonstrating PAC learnability.

Hint: The set of classifiers considered is a Dudley class. The bound you give must depend on d. Below we find a bound that does not depend on d in the realizable case, under a restriction on the width of the margin.

(b) Given $x \in \mathbb{R}^d$ and a classifier (w, b) with $w \neq 0$, let $\pi(x)$ denote the projection of x onto the hyperplane defined by $\langle w, x\rangle + b = 0$. Express $\pi(x)$ and the distance, $\|x - \pi(x)\|$, between x and the hyperplane in terms of x, w, and b.
Hint: Since w is normal to the hyperplane, $\pi(x)$ is the point in the hyperplance of the form $\pi(x) = x - cw$ for some constant c.

(c) Given a data set $Z^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ and a classifier (w, b) with $w \neq 0$, let the margin, M_i, of the ith sample point be defined by $M_i := Y_i(\langle w, X_i\rangle + b)/\|w\|$. Thus, M_i is the signed distance of X_i from the hyperplane defined by $\langle w, x\rangle + b = 0$, with the sign being positive if $Y_i = \operatorname{sgn}(\langle w, x_i\rangle + b)$ and negative otherwise. Define the margin for the whole data set by $M := \min_{i\in[n]} M_i$. Suppose that $M > 0$ for some choice of (w, b). A key idea of SVMs is to find (w, b) to maximize M, with the hope that it will lead to a better classifier for fresh samples. Show that:

$$\max_{(w,b)} M = \max \left\{ \frac{1}{\|w\|} : (w, b) \text{ subject to } Y_i(\langle w, X_i\rangle + b) \geq 1 \text{ for } i \in [n] \right\} \tag{1}$$

Hint: M for a given (w, b) is not changed if (w, b) is multiplied through by a positive scalar.

Remark: The right-hand side of (1) represents an optimization problem that is equivalent to the quadratic optimization problem (2) below.

(d) (Bound not depending on d, realizable case with lower bound on relative margin) Suppose $C_K > 0$ and $\lambda > 0$. Let $\mathcal{P}$ denote the set of all probability distributions P on $\mathsf{Z} = \mathsf{X}\times\{\pm 1\}$ such that: $P\{\sqrt{1 + \|X\|^2} \leq C_K\} = 1$, and there exists a classifier (w, b) (depending on P) such that $\|w\|^2 + b^2 \leq \lambda^2$ and $P\{Y(\langle w, X\rangle + b) \geq 1\} = 1$. These assumptions ensure that iid samples generated by P satisfy the following with probability one: $\|X_i\| \leq C_K$ for each i, and there exists (w, b) for the data points with margin M at least $1/\lambda$. Thus, the ratio of the margin to $\max_i \|X_i\|$ is greater than or equal to $\frac{1}{\lambda C_K}$. Of course, just because the data samples can be separated by a particular hyperplane doesn't necessarily mean that the hyperplane will classify fresh sample points well. Show that if $(\widehat{w_n, b_n})$ is the particular ERM classifier given by

$$(\widehat{w_n, b_n}) = \arg\min \left\{ \|w\|^2 : (w, b) \text{ subject to } Y_i(\langle w, X_i\rangle + b) \geq 1 \text{ for } i \in [n] \right\}, \tag{2}$$

then with probability at least $1 - \delta$,

$$L_P((\widehat{w_n, b_n})) \leq \frac{4\lambda C_k}{\sqrt{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{3}$$

The bound (3) does not depend on the dimension, d, of the feature space.

Hint: Bring in a Mercer kernel K, and use the ramp penalty function with unit scale parameter: $\varphi(x) = \min\{1, (1 + x)_+\}$.