ECE 543: Statistical Learning Theory

Maxim Raginsky

Homework 3

Assigned March 1; due March 8

Required reading: lecture notes of Hajek and Raginsky, Chapter 6

---

1. Let X be a compact (i.e., closed and bounded) subset of $\mathbb{R}^d$, and let $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be a Mercer kernel on X. Fix a probability distribution P supported on X, i.e., $P(\mathsf{X}) = 1$, and consider the Hilbert space $L^2(P)$ of functions $g : \mathsf{X} \to \mathbb{R}$ satisfying

$$\int_{\mathsf{X}} g^2(x) P(dx) < \infty.$$

The inner product and the corresponding norm are given by

$$\langle g, g' \rangle = \int_{\mathsf{X}} g(x) g'(x) P(dx) = \mathbf{E}[g(X) g'(X)]$$

and

$$\|g\| = \sqrt{\int_{\mathsf{X}} g^2(x) P(dx)} = \sqrt{\mathbf{E}[g^2(X)]},$$

respectively. In all cases, the expectation is w.r.t. P. The celebrated *Mercer's theorem* states the following: Consider the linear operator T_K that maps a function $g \in L^2(P)$ to a function

$$T_K g(x) := \int_{\mathsf{X}} K(x, x') g(x') P(dx') \equiv \mathbf{E}[K(x, X) g(X)], \qquad \forall x \in \mathsf{X}.$$

Then the Hilbert space $L^2(P)$ has an orthonormal basis $\{\varphi_1, \varphi_2, \ldots\}$, consisting of eigenfunctions of T_K, i.e., $\mathbf{E}[\varphi_j(X)\varphi_k(X)] = \delta_{jk}$ for all j, k, and for each j we have

$$\int_{\mathsf{X}} K(x, x') \varphi_j(x') P(dx') = \lambda_j \varphi_j(x), \qquad \forall x \in \mathsf{X}$$

(we can write this more succinctly as $T_K \varphi_j = \lambda_j \varphi_j$). Moreover, the kernel K can be represented as

$$K(x, x') = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x) \varphi_j(x'), \qquad \forall x, x' \in \mathsf{X}.$$

Armed with Mercer's theorem, prove the following two statements:

(a) Let $J := \{j \in \mathbb{N} : \lambda_j > 0\}$, and for each $j \in J$ define the function $\psi_j := \sqrt{\lambda_j}\varphi_j$. Then $\{\psi_j\}_{j\in J}$ is an orthonormal system in the RKHS $\mathscr{H}_K$, i.e., $\langle \psi_j, \psi_k \rangle_K = \delta_{jk}$ for all $j,k \in J$.
Hint: Use the reproducing property of K.

(b) Let $\mathscr{F}$ be the unit ball of $\mathscr{H}_K$, and let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. from P. Then

$$\mathbf{E}R_n(\mathscr{F}(X^n)) \leq \sqrt{\frac{1}{n}\sum_{j=1}^{\infty}\lambda_j}.$$

2. **Surrogate loss bound for a sigmoidal classifier.** Let the feature space X be a subset of $\mathbb{R}^d$. Consider the class $\mathscr{F}_R$ of functions f the form

$$f_w(x) := \tanh\left(\langle w, x\rangle\right), \tag{1}$$

where w runs over all vectors in $\mathbb{R}^d$ with $\|w\| \leq R$. Each such f induces a classifier $g_f(x) = \operatorname{sgn} f(x)$. A classifier of this kind first computes a weighted sum of the features, then passes it through a smooth nonlinear function, and then computes the sign of the resulting value. The hyperbolic tangent is an example of a *sigmoidal function* (where "sigmoidal" is a fancy term for "S-shaped" — look at the graph of $u \mapsto \tanh u$). The transformation in (1) is a simple model of a nonlinear neuron.

Let φ be a surrogate loss function satisfying the assumptions of Theorem 6.3 in the lecture notes. Let $\widehat{f}_n \in \mathscr{F}_L$ be a function generated by an arbitrary learning algorithm on the basis of an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ from an unknown probability distribution P on $\mathsf{X} \times \{-1,+1\}$. Prove that

$$L(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + 8RM_\varphi\sqrt{\frac{\mathbf{E}[\|X\|^2]}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

with probability at least $1-\delta$, where M_φ is defined in Theorem 6.3. Recall that $L(f)$ is our shorthand for the error probability $\mathbf{P}[\operatorname{sgn} f(X) \neq Y]$.

3. **AdaBoost.** In this problem, you will derive the update equations used by AdaBoost. Let $\mathscr{G}$ be a set of base classifiers mapping X to the label set $\{-1,+1\}$. The elements of $\mathscr{G}$ are called *weak learners* — think of them as being very simple. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a collection of n labeled data points, and let $w = (w_1, \ldots, w_n)$ be a probability vector, i.e., $w_i \geq 0$ for each i and $\sum_{i=1}^n w_i = 1$. You can think about w_i as the 'importance weight' attached to the ith training example. For each $g \in \mathscr{G}$, define the *weighted empirical error*

$$e_w(g) := \sum_{i=1}^n w_i \mathbf{1}_{\{y_i \neq g(x_i)\}}. \tag{2}$$

(a) Show that minimizing (2) over $g \in \mathscr{G}$ is equivalent to finding

$$\widehat{g}_w := \operatorname*{arg\,min}_{g\in\mathscr{G}} \sum_{i=1}^n w_i \exp\left(-Y_i \alpha g(X_i)\right) \tag{3}$$

for a fixed but arbitrary $\alpha > 0$.
Hint: Remember that the functions in $\mathscr{G}$ take values in $\{-1,+1\}$.

(b) Let (X, Y) be a random couple with values in $\mathsf{X} \times \{-1, +1\}$. For a classifier $g : \mathsf{X} \to \{-1, +1\}$, consider the expected surrogate loss $\mathbf{E}[e^{-Y\alpha g(X)}]$ for some $\alpha \in \mathbb{R}$. Find the value of α that minimizes $\mathbf{E}[e^{-Y\alpha g(X)}]$. Express your answer in terms of $e := \mathbf{P}[Y \neq g(X)]$. Also, show that if $e \leq \frac{1}{2} - \gamma$ for some $\gamma > 0$, then

$$\min_{\alpha \in \mathbb{R}} \mathbf{E}[e^{-Y\alpha g(X)}] = 2\sqrt{e(1-e)} \leq \exp(-2\gamma^2).$$

(c) AdaBoost forms a classifier $\widehat{f}_n$ as a convex combination of weak learners, i.e., $\widehat{f}_n(x) = \sum_{k=1}^{K} p_k g_k(x)$, where $K \in \mathbb{N}$, $p = (p_1, \ldots, p_K)$ is a probability vector, and $g_1, \ldots, g_K \in \mathcal{G}$. Consider the empirical surrogate loss of f with $\varphi(u) = e^u$:

$$A_{\varphi,n}(f) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-Y_i \sum_{k=1}^{K} p_k g_k(X_i) \right). \tag{4}$$

AdaBoost adds new weak learners to $\widehat{f}_n$ one at a time, so, to examine one step of the algorithm, suppose $p_1, \ldots, p_{k-1}$ and $g_k, \ldots, g_{k-1}$ are already given. The idea is to use a greedy approach: select α_k and g_k to minimize $A_{\varphi,n}(f)$. Show that such minimization can be decomposed into two steps:

Step one: Find $g_k \in \mathcal{G}$ to minimize the weighted empirical probability of error, $e_{w^{(k)}}$, using weights

$$w_i^{(k)} \propto \exp\left(-Y_i \sum_{j=1}^{k-1} p_j g_j(X_i) \right), \tag{5}$$

starting with $w_1^{(1)} = \ldots = w_n^{(1)} = 1/n$. For the definition of empriical probability of error, the weights need to be normalzed to sum to one.

Step two: Find p_k.

Explain why the weights $w_1^{(k)}, \ldots, w_n^{(k)}$ given in (5) are appropriate for the first step, and then describe the choice of p_k. (You can assume $p_k > 0$, which will be true if $e_k = e_{w^{(k)}}(g_k) < 1/2$, meaning that g_k does better than random guessing.)

(d) Let $f^{(k)}$ denote the classifier constructed by the algorithm after k iterations. Show that

$$A_{\varphi,n}(f^{(k)}) = A_{\varphi,n}(f^{(k-1)}) 2\sqrt{e_k(1-e_k)}$$

for $k \geq 1$.

Hence, if the weak learner has the guarantee $e_k \leq \frac{1}{2} - \gamma$ for all k (i.e., we can always find a weak learner with weighted empirical 0-1 loss less than or equal to $\frac{1}{2} - \gamma$ for any weighted data sample), then by induction on k, $A_{\varphi,n}(f^{(k)}) \leq \exp(-2k\gamma^2)$. That is, the empirical surrogate risk, and hence also the 0-1 risk, converges to zero.