Mehryar Mohri
Advanced Machine Learning 2018
Courant Institute of Mathematical Sciences
Homework assignment 2
April 30, 2018
Due: May 14, 2018

A. Learning kernels

In this problem, we will derive an alternative guarantee for learning kernels. We will use the notation adopted in class.

1. Using the results presented in class, prove the following equality:

$$\widehat{\mathfrak{R}}_S(H_1) = \frac{1}{m}\,\mathrm{E}\left[\max_{k\in[p]} \sqrt{\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}}\right].$$

2. Compute $\mathrm{E}[\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}]$, for any $k \in [p]$.

3. Prove the following inequality for the empirical Rademacher complexity of H_1:

$$\widehat{\mathfrak{R}}_S(H_1) \leq \frac{1}{m}\sqrt{\max_k \mathrm{Tr}[\mathbf{K}_k] + m\lambda_{\max}\sqrt{\frac{\log p}{2}}},$$

   where $\lambda_{\max}$ is the largest eigenvalue of a matrix $\mathbf{K}_k$, $k \in [p]$ (*hint*: you can use Jensen's inequality and the proof technique in Massart's Lemma).

B. Model selection and convex surrogates

In class, we proved that the SRM technique benefits from very favorable learning guarantees. However, SRM requires solving multiple ERM problems, which in general are NP-hard problems. Here, we will discuss guarantees for using a convex surrogate loss instead of the original binary loss.

The hypotheses we consider are real-valued functions $h\colon X \to \mathbb{R}$. The sign of h defines a binary classifier $f_h\colon X \to \{-1, +1\}$ defined for all $x \in X$ by $f_h(x) = 1_{h(x)\geq 0} - 1_{h(x)<0}$. The loss or error of h at point $(x, y) \in X \times \{-1, +1\}$ is defined as the binary classification error of f_h:

$$1_{f_h(x)\neq y} = 1_{yh(x)<0} + 1_{h(x)=0 \wedge y=-1} \leq 1_{yh(x)\leq 0}.$$

1. Show that, for any h, the generalization error of h can be expressed as follows, where $\eta(x) = \mathbb{P}[y = +1|x]$ and where D_X denote the marginal distribution over X:

$$R(h) = \mathop{\mathrm{E}}_{x\sim D_X}\left[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)\geq 0}\right].$$

2. Show that the Bayes classifier can be induced by h^* defined for all $x \in X$ by $h^*(x) = \eta(x) - \frac{1}{2}$. We will denote by R^* the Bayes error.

3. Prove that the following equality holds for the excess error of any hypothesis $h\colon X \to \mathbb{R}$:

$$R(h) - R^* = 2 \mathop{\mathrm{E}}_{x\sim D_X}\left[|h^*(x)|\, 1_{h(x)h^*(x)\leq 0}\right].$$

4. Let $\Phi\colon \mathbb{R} \to \mathbb{R}$ be a strictly convex and non-decreasing function so that for any $u \in \mathbb{R}$, $1_{u\leq 0} \leq \Phi(-u)$. For any h, define $\mathcal{L}_\Phi(h) = \mathrm{E}_{(x,y)\sim\mathcal{D}}\left[\Phi(-yh(x))\right]$. Show that $\mathcal{L}_\Phi(h) = \mathrm{E}_{x\sim\mathcal{D}_X}[L_\Phi(x, h(x))]$, where $L_\Phi(x,u) = \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$.

5. Let h^*_Φ be defined by $h^*_\Phi(x) = \mathrm{argmin}_{u\in[-\infty,+\infty]} L_\Phi(x,u)$. Prove that $f_{h^*_\Phi}$ is the Bayes classifier.

6. Assume that there exists $s \geq 1$ and $c > 0$ such that the following holds for all $x \in X$: $|h^*(x)|^s = \left|\eta(x) - \frac{1}{2}\right|^s \leq c^s\left[L_\Phi(x,0) - L_\Phi(x, h^*_\Phi(x))\right]$. Use Jensen's inequality to prove the following

$$R(h) - R(h^*) \leq 2c \mathop{\mathrm{E}}_{x\sim\mathcal{D}_X}\left[\left[\Phi(0) - L_\Phi(x, h^*_\Phi(x))\right] 1_{h(x)h^*(x)\leq 0}\right]^{\frac{1}{s}}.$$

7. Prove that $\Phi\big(-2h^*(x)h(x)\big) \leq L_\Phi(x, h(x))$ (*hint*: use convexity).

8. Use the previous inequalities to prove the following bound on the excess error in terms of the excess surrogate loss:

$$R(h) - R^* \leq 2c\left[\mathcal{L}_\Phi(h) - \mathcal{L}^*_\Phi\right]^{\frac{1}{s}}.$$

9. [Bonus point] Show that the assumption holds for the logistic loss with where $(\Phi(u) = \log_2(1 + e^u))$, with $s = 2$ and $c = \frac{1}{\sqrt{2}}$. Show the same for the exponential loss.