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:

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

   Solution: This follows immediately the equality at the bottom of slide 19. □

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

   Solution: By definition, for any $k \in [p]$,

$$\begin{aligned}
\operatorname{E}[\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}] &= \operatorname{E}[\sum_{i,j=1}^m \sigma_i \sigma_j [\mathbf{K}_k]_{ij}] \\
&= \sum_{i,j=1}^m [\mathbf{K}_k]_{ij} \operatorname{E}[\sigma_i \sigma_j] \\
&= \sum_{i=1}^m [\mathbf{K}_k]_{ii} = \operatorname{Tr}[\mathbf{K}_k]. \qquad (\operatorname{E}[\sigma_i \sigma_j] = 0 \text{ for } i \neq j)
\end{aligned}$$

   □

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

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

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).

Solution: By Jensen's inequality, in view of 1), we can write

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

Now, for any $t > 0$, we can write

$$\begin{aligned}
&\exp\left(t \operatorname{E}\left[\max_{k\in[p]} \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}\right]\right) \\
&\leq \operatorname{E}\left[\exp\left(t \max_{k\in[p]} \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}\right)\right] && \text{(Jensen's inequality)} \\
&= \operatorname{E}\left[\max_{k\in[p]} \exp\left(t \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}\right)\right] \\
&\leq \operatorname{E}\left[\sum_{k\in[p]} \exp\left(t \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}\right)\right] \\
&= \sum_{k\in[p]} \operatorname{E}\left[\exp\left(t \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}\right)\right] \\
&= \sum_{k\in[p]} e^{t \operatorname{Tr}[\mathbf{K}_k]} \operatorname{E}\left[\exp\left(t \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma} - t \operatorname{E}[\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}]\right)\right] && \text{(in view of 2))} \\
&\leq \sum_{k\in[p]} e^{t \operatorname{Tr}[\mathbf{K}_k]} e^{t^2 \lambda_{\max}^2 m^2/8} && \text{(Hoeffding's inequality)} \\
&\leq p\, e^{t \max_k \operatorname{Tr}[\mathbf{K}_k]} e^{t^2 \lambda_{\max}^2 m^2/8}.
\end{aligned}$$

When applying Hoeffding's inequality, we used the fact that

$$0 \leq \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma} \leq m\lambda_{\max}.$$

Taking the log of both sides gives

$$\operatorname{E}\left[\max_{k\in[p]} \boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma}\right] \leq \max_k \operatorname{Tr}[\mathbf{K}_k] + \frac{\log p}{t} + \frac{t\lambda_{\max}^2 m^2}{8}.$$

Choosing t to minimize the right-hand side ($t = \sqrt{8 \log p}/\lambda_{\max} m$) completes the proof. □

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:

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

Solution:

$$\begin{aligned} R(h) &= \mathop{\mathrm{E}}_{(x,y) \sim D} \big[1_{f_h(x) \neq y}\big] \\ &= \mathop{\mathrm{E}}_{x \sim D_X} \big[\eta(x) 1_{h(x)<0} + (1 - \eta(x)) 1_{h(x)>0} + (1 - \eta(x)) 1_{h(x)=0}\big] \\ &= \mathop{\mathrm{E}}_{x \sim D_X} \big[\eta(x) 1_{h(x)<0} + (1 - \eta(x)) 1_{h(x) \geq 0}\big]. \end{aligned}$$

□

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.

Solution: In view of (1), the Bayes classifier can be defined as assigning label $+1$ to x when $\eta(x) \geq \frac{1}{2}$, -1 and can therefore be induced by h^*. □

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

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

Solution: For any h, we can write

$$\begin{aligned}
R(h) &= \operatorname*{E}_{x\sim D_X}\Big[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)\geq 0}\Big]\\
&= \operatorname*{E}_{x\sim D_X}\Big[\eta(x)1_{h(x)<0} + (1-\eta(x))(1-1_{h(x)<0})\Big]\\
&= \operatorname*{E}_{x\sim D_X}\Big[[2\eta(x)-1]1_{h(x)<0} + (1-\eta(x))\Big]\\
&= \operatorname*{E}_{x\sim D_X}\Big[2h^*(x)1_{h(x)<0} + (1-\eta(x))\Big],
\end{aligned}$$

where we used for the last step the equation established earlier on. In view of that, for any h, the following holds:

$$\begin{aligned}
R(h) - R(h^*) &= \operatorname*{E}_{x\sim D_X}\Big[2[h^*(x)](1_{h(x)\leq 0} - 1_{h^*(x)\leq 0})\Big]\\
&= \operatorname*{E}_{x\sim D_X}\Big[2[h^*(x)]\operatorname{sgn}(h^*(x))1_{(h(x)h^*(x)\leq 0)\wedge((h(x),h^*(x))\neq(0,0))}\Big]\\
&= 2\operatorname*{E}_{x\sim D_X}\Big[|h^*(x)|\,1_{h(x)h^*(x)\leq 0}\Big],
\end{aligned}$$

which completes the proof, since $R(h^*) = R^*$. □

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 $\mathcal{L}_\Phi(h) = \operatorname{E}_{(x,y)\sim\mathcal{D}}\big[\Phi(-yh(x))\big]$. Show that $\mathcal{L}_\Phi(h) = \operatorname{E}_{x\sim\mathcal{D}_X}[L_\Phi(x,h(x))]$, where $L_\Phi(x,u) = \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$.

   Solution: By definition,

   $$\begin{aligned}
   \mathcal{L}_\Phi(h) &= \operatorname*{E}_{(x,y)\sim\mathcal{D}}\big[\Phi(-yh(x))\big]\\
   &= \operatorname*{E}_{x\sim\mathcal{D}_X}\Big[\eta(x)\Phi(-h(x)) + (1-\eta(x))\Phi(h(x))\Big].
   \end{aligned}$$

   □

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

   Solution: For a fixed $x\in X$, the minimum of $L_\Phi(x,u)$ is achieved for $-\eta(x)\Phi'(-u) + (1-\eta(x))\Phi'(u) = 0$. Since Φ' is non-decreasing, we have $u>0$ iff $\eta(x)>\frac{1}{2}$. □

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

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

Solution: By definition, we can write

$$\begin{aligned}
&R(h) - R(h^*)\\
&= \mathop{\mathrm{E}}_{x\sim\mathcal{D}_X}\Big[|2\eta(x) - 1|\, 1_{h(x)h^*(x)\leq 0}\Big]\\
&\leq \mathop{\mathrm{E}}_{x\sim\mathcal{D}_X}\Big[|2\eta(x) - 1|^s\, 1_{h(x)h^*(x)\leq 0}\Big]^{\frac{1}{s}} && \text{(Jensen's ineq.)}\\
&\leq 2c \mathop{\mathrm{E}}_{x\sim\mathcal{D}_X}\Big[\big[\Phi(0) - L_\Phi(x, h^*_\Phi(x))\big]\, 1_{h(x)h^*(x)\leq 0}\Big]^{\frac{1}{s}} && \text{(assumption).}
\end{aligned}$$

□

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

Solution: Using the convexity of Φ, we can write

$$\begin{aligned}
\Phi\big(-2h^*(x)h(x)\big) &= \Phi\big((1 - 2\eta(x))h(x)\big)\\
&= \Phi\big(\eta(x)(-h(x)) + (1 - \eta(x))h(x)\big)\\
&\leq \eta(x)\Phi((-h(x))) + (1 - \eta(x))\Phi(h(x)) = L_\Phi(x, h(x)).
\end{aligned}$$

□

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

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

Solution: Starting with the inequality previously proven, we can write

$$
\begin{aligned}
& R(h) - R(h^*) \\
& \leq 2c \mathop{\mathrm{E}}_{x \sim \mathcal{D}_X} \left[\left[\Phi(0) - L_\Phi(x, h^*_\Phi(x))\right] 1_{h(x)h^*(x) \leq 0} \right]^{\frac{1}{s}} \\
& \leq 2c \mathop{\mathrm{E}}_{x \sim \mathcal{D}_X} \left[\left[\Phi\big(-2h^*(x)h(x)\big) - L_\Phi(x, h^*_\Phi(x))\right] 1_{h(x)h^*(x) \leq 0} \right]^{\frac{1}{s}} && (\Phi \text{ non-decreasing}) \\
& \leq 2c \mathop{\mathrm{E}}_{x \sim \mathcal{D}_X} \left[\left[L_\Phi(x, h(x)) - L_\Phi(x, h^*_\Phi(x))\right] 1_{h(x)h^*(x) \leq 0} \right]^{\frac{1}{s}} && (\text{convexity ineq.}) \\
& \leq 2c \mathop{\mathrm{E}}_{x \sim \mathcal{D}_X} \left[L_\Phi(x, h(x)) - L_\Phi(x, h^*_\Phi(x)) \right]^{\frac{1}{s}}.
\end{aligned}
$$

□

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.