MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2022**

Problem Set 2

Due Date: **Thu., Oct. 27 2022, 10:59 am** *(Canvas submission)*

Instructions: There are **4 problems** in total in this problem set. The breakdown of individual scores per sub-problem are provided. Use the provided LaTeX template to typeset your report. Provide sufficient explanations in all solutions but avoid proving lecture or out-of-scope material (unless explicitly asked to).

What to submit: Submit your report **online through Canvas** by the due date/time. Submission must be a single pdf in LaTeX format.

Policies: Collaborative reports are not allowed. Even if you discuss problems with classmates, you are expected to write and submit **individual reports**.

Problem 1 [25 points] In this problem we will theoretically justify using the square loss as a proxy for minimizing the misclassification error by bounding the latter in terms of the former. Recall that the expected error of $f : \mathcal{X} \to \mathbb{R}$ for the square loss is

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 d\rho(x, y),$$

and denote by $f_\rho = \operatorname{argmin}_f \mathcal{E}(f)$ its minimizer. For a classification setting, where the output space is $\mathcal{Y} = \{-1, 1\}$, denote the misclassification error of f as

$$R(f) = \mathbb{P}\{\operatorname{sign}(f(x)) \neq y\},$$

where $\operatorname{sign}(\cdot)$ is defined as

$$\operatorname{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

and the probability is taken according to the distribution $\rho(x, y)$. Prove that

$$0 \leq R(f) - R(f_\rho) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho)},$$

by showing the following intermediate steps:

1.1 [10 points] $|R(f) - R(f_\rho)| = \int_{\mathcal{X}_f} |f_\rho(x)| d\rho(x)$, where $\mathcal{X}_f = \{x \in \mathcal{X} \mid \operatorname{sign}(f(x)) \neq \operatorname{sign}(f_\rho(x))\}$.

1.2 [8 points] $\int_{\mathcal{X}_f} |f_\rho(x)| d\rho(x) \leq \int_{\mathcal{X}_f} |f_\rho(x) - f(x)| d\rho(x) \leq \sqrt{\mathbb{E}(|f(x) - f_\rho(x)|^2)}$.

1.3 [7 points] $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \mathbb{E}(|f(x) - f_\rho(x)|^2)$.

Problem 2 [30 points] (Generalization error on finite hypotheses space) Recall Hoeffding's inequality: if $U_1, \ldots, U_n, U$ are i.i.d. real random variables with values in $[a, b]$, then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}U \geq t\right) \leq \exp\left\{-\frac{2nt^2}{(a-b)^2}\right\}$$

and

$$P\left(\mathbb{E}U - \frac{1}{n}\sum_{i=1}^{n} U_i \geq t\right) \leq \exp\left\{-\frac{2nt^2}{(a-b)^2}\right\}$$

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = [-M, M]$ for some $0 < M < \infty$ and consider a training set of n points sampled i.i.d from a fixed probability distribution P. Consider the square loss function ℓ and a hypothesis space comprised of N distinct functions, $\mathcal{F} = \{f_1, \ldots, f_N\}$ which are uniformly bounded, i.e. $\sup_{x\in\mathcal{X}} |f(x)| \leq C$ for all $f \in \mathcal{F}$. Recall that $\mathbf{L}(f) = \mathbb{E}\ell(Y, f(X))$ is the expected risk, and $\widehat{\mathbf{L}}(f)$ the empirical risk $\widehat{\mathbf{L}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$.

In this problem, we will derive learning guarantees in high probability *without* having to go through Rademacher averages.

2.1 [10pts] By applying Hoeffdings's inequality, derive an explicit bound on the probability

$$\Pr\left(\max_{f\in\mathcal{F}} \left|\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right| \geq \epsilon\right) \quad \forall \epsilon > 0. \tag{2.1}$$

2.2 [10pts] Let $\widehat{f}_n$ be the minimizer of the empirical risk on $\mathcal{F}$. Show that (2.1) implies that for any $0 < \delta \leq 1$, with probability at least $1 - \delta$, we have

$$\mathbf{L}(\widehat{f}_n) \leq \widehat{\mathbf{L}}(\widehat{f}_n) + \epsilon(n, N, \delta) \tag{2.2}$$

for some suitable function $\epsilon(n, N, \delta)$.

2.3 [10pts] Let $f_\mathcal{F}$ be the minimizer of the expected risk on $\mathcal{F}$. Show that (2.1) also implies that with probability at least $1 - \delta$ we have

$$\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_\mathcal{F}) \leq 2\epsilon(n, N, \delta). \tag{2.3}$$

Hint: add and subtract a few terms as we did in class and study the two difference-components of the expression individually.

Problem 3 [30 points] One of the properties of learning algorithms that we saw in class was stability, which ensures that the returned hypothesis does not change by much if the training dataset is perturbed in some fashion. There are a number of different definitions of stability, and in this problem we will explore the relationship between a few of them. In

this problem we will assume that the learning algorithms are *symmetric* in the sense that the order in which the samples are presented to it does not matter.

Let $f_{\mathcal{S}} \in \mathcal{F}$ be the hypothesis returned by a learning algorithm using the dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{z_i\}_{i=1}^n \in \mathcal{Z}^n = (\mathcal{X} \times \mathcal{Y})^n$. Let $V : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ be the loss function, that is, $V(f, z)$ is the loss incurred by the hypothesis $f \in \mathcal{F}$ on the sample $z \in \mathcal{Z}$. Finally let $\mathcal{S}_i$ be the *leave-one-out* dataset which drops z_i from $\mathcal{S}$, and $\mathcal{S}_{i,z}$ be the *replace-one* dataset which replaces z_i in $\mathcal{S}$ with z.
A learning algorithm has *uniform leave-one-out stability* β if $\forall \mathcal{S} \in \mathcal{Z}^n$, $\forall i$, and $\forall z \in \mathcal{Z}$ we have:

$$|V(f_{\mathcal{S}}, z) - V(f_{\mathcal{S}_i}, z)| \leq \beta$$

A learning algorithm has *uniform replace-one stability* β if $\forall \mathcal{S} \in \mathcal{Z}^n$, $\forall i$, and $\forall z, z' \in \mathcal{Z}$ we have:

$$|V(f_{\mathcal{S}}, z') - V(f_{\mathcal{S}_{i,z}}, z')| \leq \beta$$

Problem 3.1 [5 points] Show that a learning algorithm that has uniform leave-one-out stability β is also uniformly replace-one stable. What is its replace one stability factor?

Hint: the triangle inequality might help here.

Problem 3.2 [8 points] Now let us consider learning algorithms that take inputs that are weighted by a probability distribution p over $\mathcal{Z}$ with finite support and return a hypothesis f_p.
A learning algorithm is said to have L_1-stability λ if $\forall z \in \mathcal{Z}$ and any two distributions p, q

$$|V(f_p, z) - V(f_q, z)| \leq \lambda ||p - q||_1$$

Where $||p - q||_1 = \sum_{z \in \mathcal{Z}} |p(z) - q(z)|$ is the L_1 distance between p and q.

Show that if a learning algorithm has L_1 stability λ, then it also has uniform leave-one-out stability $2\lambda/n$.

Hint: Note that by taking a distribution that places equal mass on the points $z_i \in \mathcal{S}$ and zero everywhere else we get $f_p = f_{\mathcal{S}}$. Think about what distribution results in $f_{\mathcal{S}_i}$.

Problem 3.3 [10 points] Show that a learning algorithm has L_1 stability λ if and only if it has uniform replace-one stability $2\lambda/n$

Hint: One direction of this problem is easy to show. For the other direction, to handle two arbitrary distributions, consider a sequence of distributions that only differ at one point.

Problem 3.4 [7 points] We will now consider a specific learning algorithm - the k-nearest neighbor rule for classification. For a binary classification problem, the algorithm assigns a label ± 1 to any point $\mathbf{x} \in \mathcal{X}$ based on a majority vote of its k-nearest neighbors in the training dataset. We also take the $0-1$ loss function $V(f, z) = \mathbf{1}\{f(\mathbf{x}) \neq y\}$. Show that k-nearest neighbors has *expected leave-one-out stability* $\beta = \frac{k}{n}$. That is,

$$\mathbb{E}_{\mathcal{S},z}\left[|V(f_{\mathcal{S}},z)-V(f_{\mathcal{S}_i},z)|\right]\leq\beta=\frac{k}{n}$$

Hint: first relate the LHS of the above equation to $\Pr\left(f_{\mathcal{S}}(\mathbf{x})\neq f_{\mathcal{S}_i}(\mathbf{x})\right)$*, next argue by symmetry.*

Problem 4 [15 points] [Let $\mathcal{X}=\mathbb{R}^d$, $\mathcal{Y}=\mathbb{R}$, $\mathcal{F}\subset\{f:\mathcal{X}\to\mathcal{Y}\}$ and $X=\{x_i\}_{i=1}^n\subset\mathcal{X}$. The (empirical) Rademacher complexity of $\mathcal{F}$ is defined as follows:

$$\mathcal{R}_X(\mathcal{F})=\mathbb{E}_{\sigma_1\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i f(x_i)\right],$$

where $\sigma_1,\ldots,\sigma_n$ are i.i.d. uniformly distributed random variables over $\{\pm1\}$.

Problem 4.1 [7 points] Let $X=\{x_i\}_{i=1}^n\subset\{x\mid\|x\|\leq B\}\subset\mathcal{X}$. Suppose we have a set of bounded linear functions $\mathcal{F}_W=\{f_w(x)=\langle x,w\rangle\mid\|w\|\leq W\}$. Prove that $\mathcal{R}_X(\mathcal{F}_W)\leq\frac{BW}{\sqrt{n}}$ and that $\mathcal{R}_X(\mathcal{F}_\infty)=\infty$.

SAME AS IN NOTES

Hint: use the Cauchy-Schwarz inequality and then Jensen's inequality.

Problem 4.2 [8 points] Let $\mathcal{F}_1,\mathcal{F}_2\subset\{f:\mathcal{X}\to\mathcal{Y}\}$ and $\mathcal{F}_1+\mathcal{F}_2=\{f_1+f_2\mid f_1\in\mathcal{F}_1,f_2\in\mathcal{F}_2\}$. Prove that $\mathcal{R}_X(\mathcal{F}_1+\mathcal{F}_2)=\mathcal{R}_X(\mathcal{F}_1)+\mathcal{R}_X(\mathcal{F}_2)$.