Northwestern | McCORMICK SCHOOL OF ENGINEERING

Practice Final - Statistical Learning

Spring 2024 - Yiping

Name: ____________________ **NetID:** ____________________

Please check ☑ your Professor's name:

□ Professor Yiping Lu

While you wait, please read and check ☑ the following boxes:

□ Unless I have extra time with the Moses Center, the time limit is **100 minutes**.

□ I wrote my name and NetID (e.g. ab1234) at the top of this page.

□ I will not detach any pages, especially not the scratch pages at the end.

□ Except for multiple choice questions, I will show my work.

□ If I need more space for an exercise, I will make a note and continue on one of the scratch pages.

□ If I am caught in violation of academic integrity, including but not limited to peaking at another student's work, allowing another student to copy from my work, or speaking with another student, or using unauthorized resources, I will be asked to leave the exam and get a zero.

Do not start the exam until you are permitted to.

Exercise I

Recall that the Rademacher complexity of a class of functions $\mathscr{F}$ is defined as

$$R_n(\mathscr{F}) = \mathbb{E}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right],$$

where $Z_1, \dots, Z_n$ are drawn i.i.d. from some distribution p^* and $\sigma_1, \dots, \sigma_n$ are Rademacher variables drawn i.i.d. from $\{-1, 1\}$ with equal probability of $+1$ and -1.

(a) Let $f : \mathscr{X} \to \mathbb{R}$ be a function, and let $\mathscr{F} := \{-f, f\}$ be a function class containing only two functions. Upper bound $R_n(\mathscr{F})$ using a function of n and $\mathbb{E}[f(X)^2]$.

(b) In applications such as natural language processing, we often have sparse feature vectors. Suppose that $x \in \{0,1\}^d$ has only k non-zero entries. For example, in document classification, one feature might be “$x_{17} = 1$ iff the document contains the word *cat*."

Define the class of linear functions whose coefficients have bounded L_∞ norm:

$$\mathscr{F} = \{x \mapsto w \cdot x : \|w\|_\infty \leq B\}.$$

Compute an upper bound on the Rademacher complexity $R_n(\mathscr{F})$. Express your answer as a function of B, k, d, and n. Note that this allows us to effectively control the complexity of learning using L_∞ regularization.

(c) Consider a prediction problem from $x \in \mathbb{R}$ to $y \in \{0, \dots, k\}$. For every parameter vector $\theta \in \mathbb{R}^k$, define the prediction function $h_\theta(x) = \sum_{i=1}^{k} \mathbb{I}\{x \geq \theta_i\}$ (monotonically increasing piecewise constant functions). Define the loss function to be $\ell(y, p) = |y - p|$, yielding the following loss class:

$$\mathscr{A} = \{(x, y) \mapsto \ell(y, h_\theta(x)) : \theta \in \mathbb{R}^k\}.$$

Compute an upper bound on the Rademacher complexity of $\mathscr{A}$.

(d) Let $\mathscr{F}$ be the class of all continuous functions $f : [0,1] \to [0,1]$ with at most k local maxima. Find an upper bound of the Rademacher complexity of $\mathscr{F}$.

(e) Let X_i be independent with support $\{x \in \mathbb{R}^d : \|x\|_2 \leq M\}$. Let $\mathscr{F}$ be functions of the form $x \mapsto \langle\theta, x\rangle$ for $\theta \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq r\}$. Give an upper bound on $R_n(\mathscr{F})$.

(f) Let X_i be independent with support $\{x \in \mathbb{R}^d : \|x\|_\infty \leq M\}$. Let $\mathscr{F}$ be functions of the form $x \mapsto \langle\theta, x\rangle$ for $\theta \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$. Give an upper bound on $R_n(\mathscr{F})$.

(g) Suppose k is a bounded kernel with $\sup_x \sqrt{k(x,x)} = B < \infty$ and let $\mathscr{F}$ be its RKHS. Let $M > 0$ be fixed. Then for any $S = (X_1, \dots, X_n)$,

$$\widehat{\mathscr{R}}_S(B_k(M)) \leq \frac{MB}{\sqrt{n}}$$

where $B_k(M) = \{f \in \mathscr{F} \mid \|f\|_{\mathscr{F}} \leq M\}$.

Blank page ofr Exercise I.

Exercise II

(a) For function class

$$\mathscr{F} = \{f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ is 1-Lipschitz and convex}\},$$

show that $\log N(\epsilon, \mathscr{F}, \|\cdot\|_\infty) \lesssim \sqrt{\frac{1}{\epsilon}}$.

(b) For function class

$$\mathscr{F} = \{f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ is } L\text{-Lipschitz}\},$$

show that $\log N(\epsilon, \mathscr{F}, \|\cdot\|_\infty) \lesssim \frac{L}{\epsilon}$.

(c) For function class

$$\mathscr{F}_q = \{f_\theta(\cdot) = \langle \theta, \cdot \rangle : \theta \in B_q^d(R)\} \quad \text{with} B_q^d(R) := \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^{d} |\theta_j|^q \leq R \right\},$$

(Here the dimension d is allowed to be larger than the sample size n.) Consider $q = 1$ (i.e., ℓ_1-constrained linear regression, or Lasso) and assume that the columns of X are normalized to have ℓ_2 norm bounded by $\sqrt{n}$. Show that

$$\log N(s, B_1^d(R), \|\cdot\|_2) \lesssim R^2 \left(\frac{1}{s}\right)^2 \log d.$$

(d) Show the covering number estimation for Sobolev Ellipsoid.

(e) Using the Covering Number Bound to show the bound on Rademacher Complexity

Blank page ofr Exercise II.

Exercise III

Another view of RKHS's is in terms of **feature maps**. Let $\mathscr{F}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathscr{F}}$, which we call the feature space. It is a theorem (known as Mercer's theorem) that if k is a positive definite kernel, there is a Hilbert space $\mathscr{F}$ and function $\varphi : \mathscr{X} \to \mathscr{F}$ such that

$$k(x,z) = \langle \varphi(x), \varphi(z) \rangle_{\mathscr{F}}.$$

Of course, by our construction above, given a PSD function (kernel) k and associated RKHS $\mathscr{H}$, we can always take $\varphi(x) = k(\cdot, x)$ and $\mathscr{F} = \mathscr{H}$ directly.

(a) Let $\varphi : \mathscr{X} \to \mathscr{F}$ for a Hilbert (feature) space $\mathscr{F}$. Show that $k(x,z) = \langle \varphi(x), \varphi(z) \rangle_{\mathscr{F}}$ is a valid kernel.

(b) Consider the Gaussian or Radial Basis Function (RBF), defined on $\mathbb{R}^d \times \mathbb{R}^d$ by

$$k(x,z) = \exp\left(-\frac{1}{2}\|x - z\|_2^2\right).$$

Exhibit a function $\varphi : \mathbb{R} \to \mathbb{C}$ and distribution P on $\mathbb{R}^d$ such that

$$k(x,z) = \mathbb{E}_P\left[\varphi(W^\top x)^* \varphi(W^\top z)\right] \quad \text{for } W \sim P,$$

where $*$ denotes the complex conjugate. Is k a valid kernel?

(c) Consider the min function, defined on $\mathbb{R}_+$ by

$$k(x,z) = \min\{x, z\}.$$

Exhibit a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ such that

$$k(x,z) = \int_0^\infty \varphi(x,t)\varphi(z,t)\,dt.$$

Is k a valid kernel?

Blank page for Exercise III.

Exercise IV

The *maximum mean discrepancy (MMD)* between distributions $\mathbb{P}$ and $\mathbb{Q}$ is

$$d_{\mathrm{MMD}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x\sim\mathbb{Q}}[f(x)]. \tag{1}$$

Show that the MMD DRO Problem $\sup_{\mathbb{Q}:d_{\mathrm{MMD}}(\mathbb{Q},\mathbb{P})\leq\epsilon} \mathbb{E}_{x\sim\mathbb{Q}}[\ell_f(x)]$ is equivalent to $\mathbb{E}_{x\sim\mathbb{P}}[\ell_f(x)] + \epsilon\|\ell_f\|_{\mathcal{H}}$.

hint: Page 14 of `https://arxiv.org/pdf/1905.10943` and Question 1. (Hilbert Embedding of Probability) in Homework 8. This is actually a generalization of the χ^2 DRO in Homework 8.

Blank scratch page.

Blank scratch page.