STA 250 – Homework 1, due: Sun, April 20

Instructor: Dogyoon Song

Instructions: Please feel free to collaborate with other students on your homework, but you must list the names of any collaborators at the top of your homework assignment. All final write-ups must be done individually, and submissions must be made via Gradescope in a single LaTeX-produced PDF file. You don't have to submit your solutions to problems that are marked "Optional," which are rather challenging and will not be graded, but could be possibly interesting for your own learning.

Problem 1.

Recall the definition of Bayes predictor $f_* : \mathcal{X} \to \mathcal{Y}$ satisfying for all $x' \in \mathcal{X}$,

$$f_*(x') = \arg\min_{z \in \mathcal{Y}} \mathbb{E}\big[\ell(z, y) \mid x = x'\big].$$

Compute a Bayes predictor in the following settings:

(a) Let $\mathcal{Y} = \{-1, 1\}$ and suppose the loss function ℓ is given by

$$\ell(-1,-1) = \ell(1,1) = 0, \quad \ell(-1,1) = c_- > 0, \quad \ell(1,-1) = c_+ > 0,$$

where c_- is the cost of false negative and c_+ is the cost of false positive.

(b) Let $\mathcal{Y} = \mathbb{R}$ and consider $\ell(z, y) = |z - y|$.

Problem 2.

In class, we discussed asymptotic guarantees for empirical risk minimization. Specifically, assuming (1) θ^* is the unique minimizer of the population risk $R(\theta)$ and (2) $\nabla^2 R(\theta^*)$ is positive definite, then

1. $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, H^{-1} G H^{-1})$
2. $n\big(R(\hat{\theta}) - R(\theta^*)\big) \xrightarrow{d} \frac{1}{2}\|S\|^2$ where $S \sim \mathcal{N}(0, H^{-1/2} G H^{-1/2})$

where $G = \mathrm{Cov}\big(\nabla \ell(f_{\theta^*}(x), y)\big) = \mathbb{E}\big[\nabla \ell(f_{\theta^*}(x), y) \cdot \nabla \ell(f_{\theta^*}(x), y)^\top\big]$ and $H = \nabla^2 R(\theta^*)$.

(a) Complete the proof of Claim 1, verifying all details in each step.

(b) Prove Claim 2.

Problem 3.

(a) Let us verify some examples of sub-Gaussian random variables.

(i) Show that a Gaussian random variable with variance σ^2 is sub-Gaussian with variance proxy σ^2.

(ii) Show that any bounded random variable is sub-Gaussian, and give an upper bound on its variance proxy (as tight as possible) in terms of it support.

(b) Let X be a random variable. Prove the equivalence of the following statements:

(1) There exists $K_1 > 0$ such that

$$\Pr(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{K_1^2}\right) \quad \text{for all } t \geq 0.$$

(2) There exists $K_2 > 0$ such that

$$\left(\mathbb{E}|X|^p\right)^{1/p} \leq K_2\sqrt{p} \quad \text{for all } p \geq 1.$$

(3) There exists $K_3 > 0$ such that

$$\mathbb{E}e^{\lambda^2 X^2} \leq e^{K_3^2\lambda^2} \quad \text{for all } \lambda \in \left[-\tfrac{1}{K_3}, \tfrac{1}{K_3}\right].$$

(4) There exists $K_4 > 0$ such that $\mathbb{E}\exp\left(\frac{X^2}{K_4^2}\right) \leq 2$.

If $\mathbb{E}X = 0$, then these are also equivalent to:

(5) There exists $K_5 > 0$ such that

$$\mathbb{E}e^{\lambda X} \leq e^{K_5^2\lambda^2} \quad \text{for all } \lambda \in \mathbb{R}.$$

(c) Prove Hoeffding's inequality: if $X_1, \ldots, X_n$ are independent random variables such that $X_i \in [a_i, b_i]$ almost surely for all $i \in [n]$, then for any $t \geq 0$,

$$\Pr\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_a - a_i)^2}\right).$$

(You may assume $-a_i = b_i = \sigma$ for all $i \in [n]$, if you want to simplify.)

(d) (Optional) There are also concentration inequalities for random matrices. Prove the following.

Theorem 1 (Matrix Hoeffding; [Tro12, Theorem 1.3])) *Let $M_1, \ldots, M_n \in \mathbb{R}^{d\times d}$ be independent symmetric random matrices such that for all $i \in [n]$, (1) $\mathbb{E}[M_i] = 0$, (2) there exists $C_i \succeq 0$ such that $M_i^2 \preceq C_i^2$ almost surely. Then for all $t \geq 0$,*

$$\Pr\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n M_i\right) \geq t\right) \leq d \cdot \exp\left(-\frac{nt^2}{8\sigma^2}\right) \quad \textit{where} \quad \sigma^2 = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n C_i^2\right).$$

Problem 4.

(a) Let X be a nonnegative random variable with finite expectation. Show that $\mathbb{E}[X] = \int_0^\infty \Pr(X \geq t)dt$.

(b) Let $Y_1, \ldots, Y_n$ are zero-mean random variables (not necessarily independent) that are all sub-Gaussian with variance proxy σ^2. Prove that $\mathbb{E}[\max_{i\in[n]} Y_i] \leq \sigma\sqrt{2\log n}$.

(c) (Optional) Let $Z_1, \ldots, Z_n$ are independent Gaussian random variables with mean zero and variance σ^2. Prove that $\mathbb{E}[\max_{i\in[n]} Y_i] \geq c \cdot \sigma\sqrt{\log n}$ for some $c > 0$. ***Hint:*** *Sudakov's minoration)*

Problem 5.

(a) Prove Massart's finite lemma:

Proposition 1 (Massart's lemma) *Fix $\mathcal{D} = (z_1, \ldots, z_n)$, and let $\mathcal{G}_{\mathcal{D}} \coloneqq \big\{\big(g(z_1), \ldots, g(z_n)\big) : g \in \mathcal{G}\big\}$. If $\frac{1}{n}\|v\|_2^2 \le B^2$ for all $v \in \mathcal{G}_{\mathcal{D}}$, then*

$$\widehat{\mathrm{Rad}}_{\mathcal{D}}(\mathcal{G}) \le B\sqrt{\frac{2\log|\mathcal{G}_{\mathcal{D}}|}{n}}.$$

(b) Prove Dudley's theorem:

Theorem 2 (Dudley's theorem) *Let $\mathcal{G}$ be a family of functions from $\mathcal{Z}$ to $\mathbb{R}$ and $\mathcal{D} = (z_1, \ldots, z_n)$. Then*

$$\widehat{\mathrm{Rad}}_{\mathcal{D}}(\mathcal{G}) \le 12\int_0^\infty \sqrt{\frac{2\log N(\mathcal{G}_{\mathcal{D}}, \epsilon, d_{\mathcal{D}})}{n}}d\epsilon,$$

where $d_{\mathcal{D}}(v, v') = \frac{1}{\sqrt{n}}\|v - v'\|_2$ and $N(\mathcal{G}_{\mathcal{D}}, \epsilon, d_{\mathcal{D}})$ is the covering number of $\mathcal{G}_{\mathcal{D}}$ w.r.t. the metric $d_{\mathcal{D}}$.

(c) Let $\mathcal{F} = \{f_\theta(x) = \langle\theta, \varphi(x)\rangle, \|\theta\|_1 \le D\}$ and suppose that $\|\varphi(x_i)\|_\infty \le R$ almost surely. Prove that

$$\mathrm{Rad}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}\left[\|\Phi^\top\varepsilon\|_\infty\right] \le RD\sqrt{\frac{2\log(2d)}{n}}.$$

(d) (Optional) Let $p > 1$ and let q be such that $\frac{1}{p} + \frac{1}{q} = 1$. Define $\mathcal{F} = \{f_\theta(x) = \langle\theta, \varphi(x)\rangle, \|\theta\|_p \le D\}$ and suppose $\|\varphi(x_i)\|_q \le R$ almost surely. Prove that

$$\mathrm{Rad}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}\left[\|\Phi^\top\varepsilon\|_\infty\right] \le \frac{RD}{\sqrt{n}}\frac{1}{\sqrt{p-1}}.$$

(e) (Optional) (See [Ma22, Chapter 5.3]) Consider a two-layer neural network as follows.

- Let $\theta = (w, U)$ denote the parameters of the model, where $w \in \mathbb{R}^m$ and $U \in \mathbb{R}^{m\times d}$ with m denoting the number of hidden units.
- Consider $\mathcal{F} = \{f_\theta\}$ such that $f_\theta(x) = \langle w, \phi(Ux)\rangle = w^\top\phi(Ux)$ where $\phi(z) = \max(z, 0)$ is the ReLU activation function applied element-wisely.
- $\mathcal{D}_n = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$.

(i) Let $B_w, B_u > 0$, and consider

$$\mathcal{F} = \{f_\theta : \|w\|_2 \le B_w, \|u_i\|_2 \le B_u, \forall i \in [m]\}$$

where u_i denotes the i-th row of U. Suppose that $\mathbb{E}\big[\|x\|_2^2\big] \le C^2$. Show that

$$\mathrm{Rad}(\mathcal{F}) \le 2B_wB_uC\sqrt{\frac{m}{n}}.$$

(ii) Let $\xi(\theta) \coloneqq \sum_{j=1}^m |w_j|\|u_j\|_2$. Let $B > 0$ and consider

$$\mathcal{F} = \{f_\theta : \xi(\theta) \le B\}.$$

Show that if $\|x_i\|_2 \le C$ almost surely for all $i \in [n]$, then

$$\widehat{\mathrm{Rad}}_{\mathcal{D}_n}(\mathcal{F}) \le 2\frac{BC}{\sqrt{n}}.$$

Problem 6.

(a) Show that kernel $k(x, x') = (1 + x^\top x')^s$ corresponds to the set of all monomials $\prod_{i=1}^d x_i^{\alpha_i}$ such that $\sum_{i=1}^d \alpha_i \leq s$. What is the dimension of the resulting feature space?

(b) Let p be a probability distribution on a set $\mathcal{X}$, and $(\varphi_i)_{i\in I}$ (with I countable) be an orthonormal basis of the Hilbert space $L_2(p)$ of square-integrable functions. For a summable positive sequence $(\lambda_i)_{i\in I}$:

(i) Show that the function $k(x, x') = \sum_{i\in I} \lambda_i \varphi(x)\varphi_i(x')$ is a positive definite kernel.

(ii) Describe the associated feature space.

(c) [Bac24, Exercise 7.17]

(d) [Bac24, Exercise 7.18]

References

[Bac24] Francis Bach. *Learning Theory from First Principles.* MIT press, 2024.

[Ma22] Tengyu Ma. Stanford lecture notes for machine learning theory (CS229M/STATS214), June 2022.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.