PRACTICE EXAM

CSC2532 Winter 2024

University of Toronto

Name:

Student #:

Exam duration: **110 minutes**

Please check that your exam has **5 pages**, including this one. The total possible number of points is 100.

Read the following instructions carefully:

1. Exam is closed book and internet. You can use an optional A4 aid sheet - double-sided.
2. You must *show your work* to receive full credit.
3. The following is standard across all questions: We have a dataset of n samples $(x_i, y_i) \sim p(x, y)$ for $i = 1, 2, ..., n$, and

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell((y_i, x_i), f) \quad \text{and } f_* \coloneqq \operatorname*{argmin}_{f \in \mathcal{F}} R(f) = \mathbb{E}[\ell((y, x), f)],$$

   where ℓ is a loss function.
4. Enjoy the problems!!!

1. Warm-up: Rademacher Complexity and VC Dimension - 25pts.

1.1. *Convex-hull - 5pts.* Let $\mathcal{F} = \{f_1, f_2, ..., f_m\}$ be a finite set of functions. X-hull of $\mathcal{F}$ is defined as

$$\text{X-hull}(\mathcal{F}) = \left\{ \sum_{i=1}^{m} \alpha_i f_i : \text{ where } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{m} \alpha_i = 4 \right\}. \tag{1.1}$$

Show that $\mathfrak{R}_n(\text{X-hull}(\mathcal{F})) = 4\mathfrak{R}_n(\mathcal{F})$.

X-hull = 4 · Convex-hull

$R_n(\text{X-hull}(\mathcal{F})) = R_n(\text{Cvx-hull}(\mathcal{F})) \times 4$

$= R_n(\mathcal{F}) \times 4.$

1.2. *VC-dimension - 5pts.* Let $\mathcal{F}$ be the class of indicators of sets of the form $[a, b] \cup [c, d]$ in $\mathbb{R}$. Find the VC dimension of $\mathcal{F}$.

$n=4$ • • • • can be shattered (need to show)

$n=5$ • • • • • cannot be shattered.

1 0 1 0 1

$VC(\mathcal{F}) = 4.$

1.3. *Kernels - 5pts.* For an interval, $x = [a, b]$, define its length as $\text{len}(x) = b - a$. Show that the following is a kernel $k(x, x') = \text{len}(x \cap x') + \text{len}(x)\text{len}(x')$ Here, intersection of intervals is an interval or the empty set (which has length 0).

$k_1 = \text{len}(x)\, \text{len}(x')$ is a kernell ($x \to \text{len}\, x$ is a feature map)

$k_2 = \int 1_x(u) \cdot 1_{x'}(u)\, du \Rightarrow$ inner product. $\Rightarrow$ kernel

$\Rightarrow k_1 + k_2$ is a kernel.

1.4. *Representer- 10pts.* Let $\mathcal{F}$ be an RKHS and k be the associated kernel. For $x_1, x_2, ..., x_n$ iid from a distribution p, let $\hat{f} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i)$ and $f^* = \mathbb{E}[k(\cdot, x_1)]$, and $D := \|\hat{f} - f^*\|_{\mathcal{F}}$. Let $\hat{f}'$ and D' be defined similarly over $x_1', x_2, ..., x_n$ (only x_1 is different).

1- Prove that $D - D' \leq 2 \sup_x \sqrt{k(x, x)}/n$. 2- Show $\mathbb{E}[D] \leq \sup_x \sqrt{k(x, x)/n}$.

1- D is a norm $\Rightarrow$ $D - D' \leq \|\frac{1}{n}k(\cdot, x_1) - \frac{1}{n}k(\cdot, x_1')\|_{\mathcal{F}} = \frac{1}{n}\|k(\cdot, x_1) - k(\cdot, x_1')\|_{\mathcal{F}}$

$= \frac{1}{n}\left\{\|k(\cdot, x_1)\|_{\mathcal{F}} + \|k(\cdot, x_1')\|_{\mathcal{F}}\right\}$

$\leq \frac{2}{n} \sup_x \sqrt{k(x,x)}$

2- $\mathbb{E}D \leq \mathbb{E}[D^2]^{1/2}$ (Jensen)

$= \mathbb{E}[\|\hat{f} - f_*\|_{\mathcal{F}}^2]^{1/2}$

$= \mathbb{E}[\|\frac{1}{n}\sum_i k(x_i, \cdot) - f_*\|_{\mathcal{F}}^2]^{1/2}$

$$= \left(\frac{1}{n^2} \sum_i \mathbb{E} \| k(x_i, \cdot) - f_* \|_{\mathcal{F}}^2 \right)^{1/2} = \left(\frac{1}{n} \mathbb{E} \| k(x_1, \cdot) - f_* \|_{\mathcal{F}}^2 \right)^{1/2}$$

$$= \frac{1}{\sqrt{n}} \mathbb{E} \left[\| k(x, \cdot) \|_{\mathcal{F}}^2 - \| f_* \|^2 \right]^{1/2} \leq \frac{1}{\sqrt{n}} \sup_x k(x, x)^{1/2}.$$

2. Expected Excess Risk - 25pts. In class, we mostly focused on giving generalization guarantees in high probability. For example, we showed that, with probability at least $1 - \delta$, excess risk satisfies

$$R(\hat{f}) - R(f_*) \leq 4\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{2\log(1/\delta)}{n}}, \tag{2.1}$$

where $\mathcal{G} = \{(y, x) \to \ell((y, x), f) \ : \ \forall f \in \mathcal{F}\}$.

In this question, we will prove a generalization bound *in expectation*. Steps are essentially the same, though proof is simplified.

1. *[10pts]* Show that expected excess risk can be upper bounded by the supremum of the empirical process. E.g., show

$$\mathbb{E}\Big[R(\hat{f}) - R(f_*)\Big] \leq \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\Big] + \mathbb{E}\Big[\sup_{f\in\mathcal{F}} R(f) - \hat{R}(f)\Big]. \tag{2.2}$$

$$R(\hat{f}) - R(f_*) = R(\hat{f}) - \hat{R}(\hat{f}) + \underbrace{\hat{R}(\hat{f}) - \hat{R}(f_*)}_{\leq 0} + \hat{R}(f_*) - R(f_*)$$

$$\leq \sup \hat{R}(f) - R(f) + \sup R(f) - \hat{R}(f)$$

Take expectations

2. *[10pts]* Show that the right hand side of the above inequality can be upper bounded with Rademacher complexity of $\mathcal{G}$.

In lecture, we proved $\mathbb{E} \sup \hat{R}(f) - R(f) \leq 2\mathcal{R}(\mathcal{G})$ using symmetrization.

3. *[5pts]* Finally, conclude that the expected excess risk can be upper bounded by the Rademacher complexity of $\mathcal{G}$ times a constant which you should compute explicitly. Which crucial assumption on loss is missing, and why?

$$\Rightarrow \quad \mathbb{E}\left[R(\hat{f}) - R(f_*)\right] \leq 4\mathcal{R}(\mathcal{G})$$

We don't need loss to be bdd since concentration argument.

3. KL and Identifiability - 25 pts. Given two probability distributions $p(x)$ and $q(x)$ fully supported on $\mathbb{R}^d$ ($p(x) > 0$ and $q(x) > 0$ for all $x \in \mathbb{R}^d$), KL divergence is defined as

$$\mathrm{KL}(p||q) = \mathbb{E}_p\left[\log \frac{p(x)}{q(x)}\right] = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{3.1}$$

KL divergence is not a metric since it doesn't satisfy triangle inequality. However, it has nice properties, and it provides a distance measure between two distributions. One property is the following:

3.1. *KL property - 10pts.* Show that $\mathrm{KL}(p||q) = 0$ if and only if $p = q$. Hint: Jensen's inequality says if ϕ is convex, then $\mathbb{E}[\phi(x)] \geq \phi(\mathbb{E}[x])$ with equality if and only if x is constant or ϕ is linear.

$p = q \Rightarrow KL = 0$ trivial

$$\mathbb{E}_p \log p/q = \mathbb{E}_p -\log q/p$$

$$\geq -\log \mathbb{E}_p q/p = 0 \quad (-\log \text{ is cvx})$$

$KL = 0 \Rightarrow$ equality holds $\Leftrightarrow \frac{q}{p}$ is constant or $-\log$ linear.

3.2. *Identifiability in parametric families - 15pts.* Consider the parameteric family where

$$y|x \sim p_{\theta_*}(y|x) \quad \text{and} \quad x \sim p(x),$$

with $\theta_* \in \mathbb{R}^m$ is the true parameter. Under the identifiability assumption that $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$, show that the true parameter is the unique global minimizer of the population risk in the MLE setup where the loss is $\ell(\theta, (y, x)) = -\log p_\theta(y|x)$, i.e. prove

$$\theta_* = \underset{\theta \in \mathbb{R}^m}{\mathrm{argmin}}\, R(\theta) := \mathbb{E}[-\log p_\theta(y|x)] \tag{3.2}$$

where expectation is over the true distribution $(x, y) \sim p_{\theta_*}(y|x)p(x)$. Hint: Consider the quantity $R(\theta) - R(\theta_*)$ for $\theta \neq \theta_*$.

$\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$.

$$R(\theta) - R(\theta_*) = \mathbb{E}_{\theta_*} -\log P_\theta(y|x) + \log P_{\theta_*}(y|x)$$

$$= \mathbb{E}_x \mathbb{E}_{P_{\theta_*}(y|x)} - \log \frac{P_\theta(y|x)}{P_{\theta_*}(y|x)} = \mathbb{E}_x KL(P_{\theta_*}(\cdot|x) || P_\theta(\cdot|x))$$

> 0 unless $P_{\theta_*} = P_\theta$ by the assump.

4. Countable Function Class - 25 pts. Let $\mathcal{F} = \{f_1, f_2, ...\}$ be a countable set of functions with infinite size $|\mathcal{F}| = \infty$, and loss evaluated for each function satisfies

$$0 \leq \ell((x, y), f_i) \leq \frac{B}{i^\beta},$$

for some $\beta > 0$, a bound decaying with function's index.

For what values of β does this class achieve generalization? In your bounds, you should compute all constants explicitly.

Need $\sum_{f \in \mathcal{F}} \mathbb{P}\left(|\hat{R}(f) - R(f)| \geq \varepsilon \right) < \infty$. (1st lecture)

$$\leq \sum_i e^{-n\varepsilon^2 / \frac{B^2}{i^{2\beta}}} = \sum_{i=1}^{\infty} \underbrace{e^{-\frac{n\varepsilon^2}{B^2}}}_{\rho} \cdot {}^{i^{2\beta}}$$

$$= \sum_i \rho^{i^{2\beta}} \leq \sum_i \rho^i \quad \text{for } \beta \geq \frac{1}{2}$$

$$< \infty$$

! Doesn't say what happens for $\beta < 1/2$.