EE 660: Machine Learning II Fall 2024
Tuesday, October 8

Midterm Exam

Name:

USC ID:

This is an open book exam: you may use any books, lecture notes, personal notes, homework solutions, previous midterm solutions, etc. You are also allowed to use any electronic copies on your laptop and/or tablet (cell phones and other devices are prohibited). **You are not permitted to connect to the internet or communicate with anyone in anyway during the exam. Any violations of this policy during the exam will be considered cheating.**

The weight of each sub-problem is indicated on the question. The last few pages of the exam contain extra blank pages, should you run out of space when writing your answers. If you need to utilize the extra pages, please indicate so very clearly when answering the questions.

Q1	/10
Q2	/20
Q3	/25
Q4	/25
Total	/80

(Q1). True or False questions (10 points)

Circle either true (T) or false (F). No justification is needed for your answer.

(a) T / [✓ F] The sub-Gaussian maximal inequality $\mathbb{E}\max_{i\in[n]} X_i \leqslant \sigma\sqrt{2\log n}$ requires that the X_i's are independent.

(b) T / [✓ F] Given a linearly separable dataset, the Perceptron and SVM solutions always coincide.

(c) [✓ T] / F Given an arbitrary dataset $\{(x_i, y_i)\}_{i=1}^n$, there exists a feature map $\Phi(x)$ with finite output dimensionality such that $\{(\Phi(x_i), y_i)\}_{i=1}^n$ is linearly separable. $\Phi(x_i) = e_i$

(d) [✓ T] / F A kernel $k(x, y)$ cannot satisfy $k(x, x) < 0$. $k(x,x) = \|x\|^2$

(e) T / [✓ F] The kernel $k(x, y) = \langle x, y\rangle^2$ is shift-invariant.

$\forall \gamma \geq -1$

(f) T / [✓ F] For any $\gamma \in \mathbb{R}$ and $y, \hat{y} \in \{\pm 1\}$, we have $\underbrace{\mathbf{1}\{y \neq \hat{y}\}}_{=1} \leqslant \mathbf{1}\{\underset{-1}{y\hat{y}} \leqslant \gamma\}$. $y = 1$, $\hat{y} = -1$ $\gamma = -100$

(g) [✓ T] / F The VC dimension of a hypothesis class is monotonic, i.e., $\mathscr{H} \subseteq \mathscr{H}'$ implies $\mathrm{VCdim}(\mathscr{H}) \leqslant \mathrm{VCdim}(\mathscr{H}')$.

(h) T / [✓ F] Consider $\mathscr{H} = \{x \mapsto \mathrm{sgn}(f(x)) \mid f \in \mathscr{F}\}$ such that $\mathrm{VCdim}(\mathscr{H}) = \infty$. Then, we must have $\mathcal{R}_n(\mathscr{H}) = \infty$ as well.

~~(i)~~ T / F Any finite hypothesis class (i.e., $|\mathscr{H}| < \infty$) is agnostically PAC learnable.

(j) T / [✓ F] Any function class $\mathscr{F}$ with a finite number of parameters must have $\mathcal{R}_n(\mathscr{F}) < \infty$.

$$R_n(\mathcal{H}) = \frac{1}{n} E\Big[\sup_{h\in\mathcal{H}} \underbrace{\sum_{i=1}^{n} \varepsilon_i h(x_i)}\Big] \leq 1$$

$$\sum_{i=1}^{n} \varepsilon_i h(x_i) \leq \sum_{i=1}^{n} |\varepsilon_i h(x_i)| \leq n$$

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d\}$$

$$\mathcal{F}_\alpha = \{x \mapsto \langle w, x \rangle \mid \|w\| \le \alpha,\ w \in \mathbb{R}^d\} \quad \alpha > 0$$

$$\mathcal{F}_\alpha \subseteq \mathcal{F}$$

$$\mathcal{R}_n(\mathcal{F}) \geq \mathcal{R}_n(\mathcal{F}_\alpha)$$

$$= \frac{1}{n} E \sup_{\|w\| \le \alpha} \sum_{i=1}^n \varepsilon_i \langle w, x_i \rangle$$

$$= \frac{1}{n} E \sup_{\|w\| \le \alpha} \langle \sum_{i=1}^n \varepsilon_i x_i, w \rangle$$

$$w^* = \alpha \frac{\sum_{i=1}^n \varepsilon_i x_i}{\|\sum_{i=1}^n \varepsilon_i x_i\|}$$

$$= \frac{\alpha}{n} \underbrace{E \|\sum_{i=1}^n \varepsilon_i x_i\|}_{>0}$$

$$\alpha \to \infty$$

$$\mathcal{R}_n(\mathcal{F}) \geq \mathcal{R}_n(\mathcal{F}_\alpha) \to \infty$$

(Q2). Random features (20 points)

Fix a kernel $k : \mathsf{X} \times \mathsf{X} \mapsto \mathbb{R}$. Suppose there exists a map $\varphi : \mathsf{X} \times \Omega \mapsto [-1, 1]$ and a distribution $p(\omega)$ over Ω such that the following holds for all $x, y \in \mathsf{X}$:

$$k(x, y) = \mathbb{E}_{\omega \sim p(\cdot)}[\varphi(x, \omega)\varphi(y, \omega)].$$

For $m \in \mathbb{N}_+$, let $\{\omega_i\}_{i=1}^m$ be m iid draws from $p(\cdot)$, and define the feature map $\Phi : \mathsf{X} \mapsto \mathbb{R}^m$:

$$\Phi(x) := \frac{1}{\sqrt{m}}(\varphi(x, \omega_1), \ldots, \varphi(x, \omega_m)).$$

Part (a). [5 points]

Show that $\mathbb{E}[\langle \Phi(x), \Phi(y) \rangle] = k(x, y)$, where the expectation is taken with respect to $\{\omega_i\}_{i=1}^m$.

$$\begin{aligned} E[\langle \Phi(x), \Phi(y) \rangle] &= E\left[\frac{1}{m}\sum_{i=1}^m \varphi(x, \omega_i)\varphi(y, \omega_i)\right] \\ &= \frac{1}{m}\sum_{i=1}^m E[\varphi(x, \omega_i)\varphi(y, \omega_i)] \\ &= \frac{1}{m}\sum_{i=1}^m k(x, y) = k(x, y) \end{aligned}$$

Part (b). [15 points]

Let $D \subset \mathsf{X}$ denote a *finite* subset of X, with cardinality $|D| < \infty$. Fix $\delta \in (0, 1)$. Show that with probability at least $1 - \delta$ (over the randomness of $\{\omega_i\}_{i=1}^m$):

$$\max_{x,y \in D} |k(x, y) - \langle \Phi(x), \Phi(y) \rangle| \leqslant \sqrt{\frac{2 \log(2|D|^2/\delta)}{m}}.$$

Note: You will receive full credit for any valid bound of the form $c_0\sqrt{\frac{\log(c_1|D|^2/\delta)}{m}}$ where c_0, c_1 are absolute constants.

Hint: Use the *two-sided* version of Hoeffding's inequality, which states that if $X_1, \ldots, X_m$ are independent random variables with $\underline{X_i \in [-a_i, a_i]}$ for $i \in [m]$, then putting $S_m := \sum_{i=1}^m X_i$,

$$\forall t > 0, \quad \mathbb{P}\{|S_m - \mathbb{E}[S_m]| \geqslant t\} \leqslant 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^m a_i^2}\right).$$

$\alpha_i = 1 \quad S_m = \sum_{i=1}^m X_m , \quad M_m = \frac{1}{m} S_m$

$$= \underline{\sum_{i=1}^m \left(\frac{X_m}{m}\right)}$$

$\frac{X_i}{m} \in [-\frac{1}{m}, \frac{1}{m}]$

$\forall t>0 \quad \mathbb{P}\{|M_m - E[M_m]| \geq t]$

$$\leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^m (\frac{1}{m})^2}\right) = 2\exp\left(-\frac{mt^2}{2}\right)$$

$$\langle \Phi(x), \Phi(y) \rangle = \frac{1}{m}\sum_{i=1}^m \underbrace{\varphi(x_i, \omega)\varphi(y_i, \omega)}_{|\cdot| \leq 1}$$

$\forall t>0$

$$\mathbb{P}\{|\langle \Phi(x), \Phi(y)\rangle - E[\overbrace{\langle \Phi(x), \Phi(y)\rangle}^{k(x,y)}]| \geq t]$$

$$\leq 2\exp\left(-\frac{mt^2}{2}\right)$$

Take union bound over all pairs $(x,y) \in D$

$$\mathbb{P}\{\max_{x,y\in D} |\langle \Phi(x), \Phi(y)\rangle - k(x,y)| \geq t\}$$

$$= \mathbb{P}\{\bigcup_{x,y\in D} \{|\langle \Phi(x), \Phi(y)\rangle - k(x,y)| \geq t\}\}$$

$$\leq \sum_{x,y\in D} \underbrace{\mathbb{P}(|\langle \Phi(x), \Phi(y)\rangle - k(x,y)| \geq t}_{})$$

$$\leq \sum_{x,y\in D} 2\exp(-mt^2/2)$$

$$= 2|D|^2 \exp(-mt^2/2) = \delta$$

(Q3). VC dimension (25 points)

Part (a). [10 points]

Let $\mathscr{H} = \{x \mapsto h_{w,b}(x) := \operatorname{sgn}(\langle x, w\rangle + b) \mid w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$ denote the set of linear classifiers with a bias term. Show that $\operatorname{VCdim}(\mathscr{H}) \geqslant d + 1$.

Need $x_{1:d+1}$ in $\mathbb{R}^d$ s.t. $\mathcal{H}$ shatters $x_{1:d+1}$

Set $x_i = e_i \quad \forall i \in \{1, 2, \cdots, d\}$

$x_{d+1} = 0$

$y \in \{\pm 1\}^{d+1}$

$w_i = y_i - y_{d+1} \quad \forall i \in \{1, 2, \cdots, d]$

$b = y_{d+1}$

For $i \in \{1, 2, \cdots, d]$ $\quad \langle \overset{e_i}{\overset{\shortparallel}{x_i}}, w \rangle$

$h_{w,b}(x_i) = sgn(y_i - y_{d+1} + y_{d+1}) = sgn(y_i) = y_i$

$i = d+1$

$h_{w,b}(x_{d+1}) = sgn(y_{d+1}) = y_{d+1}$

$\Rightarrow \mathcal{H}$ shatters $x_{1:d+1}$

Part (b). [10 points]

Let $\mathscr{H}$ be as defined in part (a). Show that $\mathrm{VCdim}(\mathscr{H}) \leqslant d+1$.

Hint: Use the fact that for $\mathscr{H}' := \{x \mapsto \mathrm{sgn}(\langle x, \bar{w} \rangle) \mid \bar{w} \in \mathbb{R}^{d+1}\}$, we have $\mathrm{VCdim}(\mathscr{H}') = d+1$.

$\forall x \in \mathbb{R}^d$, let $\bar{x} := (x, 1) \in \mathbb{R}^{d+1}$

$\forall x_{1:n} \subset \mathbb{R}^d$

$$|\mathcal{H}(x_{1:n})| = |\underline{\mathcal{H}'(\bar{x}_{1:n})}| \boxed{\leq} \tau_{\mathcal{H}'}(n) \quad (*)$$

Let $\tau_{\mathcal{H}}(n) = \sup_{x_{1:n} \subset \mathbb{R}^d} |\mathcal{H}(x_{1:n})|$

$$\tau_{\mathcal{H}'}(n) = \sup_{x'_{1:n} \subset \mathbb{R}^{d+1}} |\mathcal{H}'(x'_{1:n})|$$

Take sup on LHS of $(*)$ over $x_{1:n} \subset \mathbb{R}^d$

$$\underline{\tau_{\mathcal{H}}(n)} \leq \tau_{\mathcal{H}'}(n)$$

$$\mathrm{VCdim}(\mathcal{H}) = \sup\{n \in \mathbb{N}_+ \mid \tau_{\mathcal{H}}(n) = 2^n\}$$

$\left(l \text{ s.t. } \tau_{\mathcal{H}}(l) = 2^l > \tau_{\mathcal{H}'}(l) \right)$

$$\leq \sup\{n \in \mathbb{N}_+ \mid \tau_{\mathcal{H}'}(n) = 2^n\}$$
$$= \mathrm{VCdim}(\mathcal{H}')$$
$$= d+1$$

$$\Rightarrow \mathrm{VCdim}(\mathcal{H}) = d+1$$

Part (c). [5 points]

Show that:

$$\mathrm{VCdim}(\{x \mapsto \mathrm{sgn}(\mathbf{1}\{x \in [a, a+1]\} - 1/2) \mid a \in \mathbb{R}\}) = 2.$$

(Q4). Cumulative distribution function (CDF) estimation (25 points)

Let μ be a distribution over $\mathbb{R}$ with cumulative distribution function (CDF) $F(t) := \mathbb{P}_{x \sim \mu}\{x \leqslant t\}$. Let $x_1, \dots, x_n$ denote n iid draws from μ, and define the empirical estimate $\hat{F}_n(t)$ of $F(t)$ as:

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \leqslant t\}, \quad t \in \mathbb{R}.$$

In this problem, we will show a classic result (known as the Glivenko–Cantelli theorem):

$$\mathbb{E} \sup_{t \in \mathbb{R}} |F(t) - \hat{F}_n(t)| \leqslant c_0 \sqrt{\frac{\log(c_1 n)}{n}},$$

where c_0, c_1 are universal positive constants. Above, the expectation on the left hand side is taken with respect to $\{x_i\}_{i=1}^n$.

Part (a). [5 points]

Let $\mathscr{H} := \{x \mapsto \operatorname{sgn}(t - x) \mid t \in \mathbb{R}\}$. Show that:

$$\mathcal{R}_n(\{x \mapsto \mathbf{1}\{x \leqslant t\} \mid t \in \mathbb{R}\}) = \frac{1}{2} \mathcal{R}_n(\mathscr{H}).$$

Hint: Use the identity $\mathbf{1}\{x \leqslant t\} = \frac{\operatorname{sgn}(t-x)+1}{2}$.

Part (b). [5 points]

Let $\mathscr{H}$ be as defined in part (a). Show that $\mathcal{R}_n(\mathscr{H}) \leqslant \sqrt{\frac{2\log(en)}{n}}$.

Note: You will receive full credit for any valid bound of the form $c_0\sqrt{\frac{\log(c_1 n)}{n}}$ where c_0, c_1 are absolute constants.

Part (c). [15 points]

Recall the following alternative definition of Rademacher complexity from Homework 2:

$$\bar{\mathcal{R}}_n(\mathscr{F}) := \mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right|.$$

One can show that this definition of Rademacher gives rise to the following upper bound:

$$\mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E}[f(x_i)] - f(x_i)) \right| \leqslant 2\bar{\mathcal{R}}_n(\mathscr{F}).$$

Use the above inequality, which you may take for granted, to show that:

$$\mathbb{E} \sup_{t \in \mathbb{R}} |F(t) - \hat{F}_n(t)| \leqslant 4\sqrt{\frac{2\log(en)}{n}}$$

Note: You will receive full credit for any valid bound of the form $c_0\sqrt{\frac{\log(c_1 n)}{n}}$ where c_0, c_1 are absolute constants.

Hint: Recall from Homework 2 that if $|f| \leqslant 1$ for all $f \in \mathscr{F}$, then $\bar{\mathcal{R}}_n(\mathscr{F}) \leqslant 2\mathcal{R}_n(\mathscr{F}) + 1/\sqrt{n}$.