Midterm Exam [Sample Exam, showing format] APPM 4490/5490 Spr 2022 Theoretical ML

Instructor: Prof. Becker

Name: ____________________

Instructions All logarithms are natural logarithms unless otherwise noted.
Your proofs may use any major result discussed in class (if you are unsure, please ask).
Total points possible: $\boxed{50.}$ Note that points are *not* necessarily distributed according to difficulty.

Problem 1: **True/False** (20 points, equally weighted; no justification needed). Answer "True" if the statement is always true, otherwise answer "False". **You may choose *one* of these to skip** (please indicate which one you do not want graded). **Please write directly on the test**.

1) Sample question . . . . ______
2) $\vdots$
3) Sample question . . . . ______

Problem 2: (10 points) Sample question

Problem 3: (20 points, 5 points each)

1) Subquestion 1
2) Subquestion 2
3) Subquestion 3
4) Subquestion 4

Problem 4: (? points) Possibly one more question

Problem 5: **Bonus** (10 extra credit points) (possibly)

Cheat sheet

$(\forall x \in \mathbb{R})\ 1 - x \leq e^{-x}$.

Sauer: If $\mathrm{VCdim}(\mathcal{H}) = d$, then

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$$

For binary classification (Thm. 6.11 [Shalev-Shwartz]),

$$(\forall h \in \mathcal{H})\, |\mathcal{L}_{\mathcal{D}}(h) - \widehat{L}_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}.$$

Fund. Thm. Stat. Learning ($0-1$ loss, $\mathcal{Y} = \{0, 1\}$), if $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, then $\mathcal{H}$ is PAC-learnable with sample complexity (for $0 < \delta < 1/4$)

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{8}{\epsilon}\left(2d \log\left(\frac{16e}{\epsilon}\right) + \log\left(\frac{2}{\delta}\right)\right)$$

(c.f. Thm 6.8 part (3) + Thm 28.3 + Thm 28.4 in [Shalev-Shwartz]), and $\mathcal{H}$ is agnostic PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq C\frac{d + \log(1/\delta)}{\epsilon^2}$$

for some constant C (Thm. 6.8 part (2)), or a slightly worse bound but with explicit constants

$$m_{\mathcal{H}}(\epsilon, \delta) \leq 4\frac{\left(8\log\left(\frac{64d}{\epsilon^2}\right) + r\log\left(\frac{e}{d}\right)\right)d + \log(4/\delta)}{\epsilon^2}$$

(section 28.1 [Shalev-Shwartz]).

For binary classification with $0-1$ loss, if $\mathcal{H}$ is a family of functions with $\{\pm 1\}$ output, then with probability at least $1 - \delta$ (Corollary 3.8 [Mohri]),

$$L_{\mathcal{D}}(h) \leq \widehat{L}_S(h) + \begin{cases} \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}} \\ \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \end{cases}$$

and $\mathfrak{R}_m(\mathcal{H}) \leq \sqrt{2m^{-1}\log(\tau_{\mathcal{H}}(m))}$.

If $c \in \mathbb{R}, a_0 \in \mathbb{R}^d$ then $\widehat{\mathfrak{R}}(cA + a_0) \leq |c|\widehat{\mathfrak{R}}(A)\ \forall A \subset \mathbb{R}^d$.

Hoeffding: If $(\theta_1, \ldots, \theta_m)$ are iid r.v., and $\mathbb{E}[\theta_i] = \mu$ and $\theta_i \in [a, b]$, then $\forall \epsilon > 0$

$$\mathbb{P}[\hat{\mu}_m > \mu + \epsilon] \leq \exp\left(-2m\epsilon^2/(b-a)^2\right)$$
$$\mathbb{P}[\hat{\mu}_m < \mu - \epsilon] \leq \exp\left(-2m\epsilon^2/(b-a)^2\right)$$

where $\hat{\mu}_m \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m \theta_i$ is the sample mean.

Lemma A.2: If $a \geq 1, b > 0$ then

$$(x \geq 4a\log(2a) + 2b) \implies (x \geq a\log(x) + b)$$

and (this version isn't in the book)

$$(x > 4a\log(2a) + 2b) \implies (x > a\log(x) + b)$$