ECE 2372 **Name:** ____________________
Midterm Exam
Time Limit: 120 Minutes

- Write your final answers in the boxes provided.
- Write down any assumptions. Show all work, you will not get credit if there isn't work to back it up.
- Your exams will be electronically scanned prior to grading. Your writing must be legible, and dark enough for the scanner to detect, for your answers to be graded.
- There are total of 5 problems in this midterm exam.
- The exam is worth 100 points. Each question will indicate how many points it is worth. In multi-part questions, each part will be weighted equally.
- This exam document has 8 pages. Notice last two pages are provided as a extra workspace.
- All work should be performed on the exam paper itself. If more space is needed, use the backs of the pages.
- This exam will be conducted under the rules and guidelines of the University of Pittsburgh's Honor Code and no cheating will be tolerated.
- Please double check and make sure that your submission contains everything you intend to submit. You will only be graded for what you submit. **Any submission made after 5:15pm on March 16 will not be graded**.

1. (20 points) (Short Answers. Answer each problem (in words) using $1-2$ short sentences.) (4 points each)

   (a) Describe the difference between Hoeffding's inequality (as we have stated and used it in this course) and the VC generalization bound.

   (b) How does Tikhonov regularization change the standard least-squares regression solution when $\Gamma = \sqrt{\lambda} I$ (where $\lambda > 0$)

   (c) In what way is the decision boundary found by maximum margin classifier "optimal"?

   (d) What does the "kernel trick" let us do and why is this useful?

2. (20 points) Consider a binary classification problem involving a single (scalar) feature x and suppose that $X|Y=0$ and $X|Y=1$ are continuous random variables with densities given by

$$g_0(x) = \frac{1}{2}e^{-|x|}$$
$$g_1(x) = \frac{1}{2}e^{-|x-1|}$$

respectively. Furthermore, suppose that $P[Y=0] = P[Y=1] = \frac{1}{2}$. Derive the optimal classification rule (in terms of minimizing the probability of error) and calculate the Bayes risk for this classification problem.

3. (20 points) Consider classifiers defined by positive circles in $\mathbb{R}^2$, i.e., h such that for any $x \in \mathbb{R}^2$,

$$h(x) = \begin{cases} +1 & \text{if } \|x - c\| \leq r \\ -1 & \text{otherwise} \end{cases}$$

for some $c \in \mathbb{R}^2$, $r \in \mathbb{R}$. Determine the VC dimension d_{VC} of the set of all such classifiers. For whatever number you give, justify your answer by using an example to show that the VC dimension must be **at least** that big, and then also provide an argument that the VC dimension cannot be any larger.

4. (20 points) Consider a learning scenario where $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ is given by $Y = X^2$. Assume that the input variable X is drawn uniformly on the interval $[0, 1]$. Now suppose that we are given a single observation of an input-output pair, i.e., our data set is given by $\mathcal{D} = \{(x_1, x_1^2)\}$. Consider the problem of fitting a line of the form $h(x) = ax$ to this data by selecting the line that passes through our one observation.

   (a) Determine the average hypothesis $\bar{h}(X) = \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X)\right]$.

   (b) Determine the bias $\mathbb{E}_X\left[(\bar{h}(X) - X^2)^2\right]$.

   (c) Determine the variance $\mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right]\right]$.

5. (20 points) Suppose that we have two independent random variables X_1 and X_2 that are both uniform on the interval $[0, 1]$. Think of X_1 and X_2 as estimates of the probability of error of two different classifiers, h_1 and h_2. We are going to decide to use h_1 if we observe $X_1 < X_2$, and h_2 if we observe $X_2 < X_1$.

   (a) What are $\mathbb{E}[X_1]$ and $\mathbb{E}[X_2]$?

   (b) Let $X^* = \min(X_1, X_2)$. Calculate $\mathbb{P}[X^* \leq x]$ where $x \in [0, 1]$ is arbitrary.

   (c) Calculate $\mathbb{E}[X^*]$. [Hint: Remember that the probability density function for X^* is given by the derivative of the cumulative density function $F_{X^*}(x) = \mathbb{P}[X^* \leq x]$.]

   (d) Can we use X^* to predict how well the classifier we chose will perform in the future? If so, how? If not, why not?

Additional workspace:

Additional workspace: