Groundrules:

- You should solve three of the six problems below.
- You are not allowed to collaborate with others on this exam.
- You are allowed to consult the lecture notes, but no other external sources.
- You must submit your exam via Autolab.

1. PAC learning.

   (a) A $k$-DNF formula over $\{0, 1\}^n$ is a disjunction (an OR) of “terms,” where each term is an AND of up to $k$ literals (a literal is either a variable or its negation). Give a polynomial-time PAC-learning algorithm for learning the class $C_{3\text{DNF}}$ of 3-DNF formulas in the realizable case. Also give an explicit sample complexity bound (you may use $O()$ notation).

   (b) A union of 3 intervals over the real line is a Boolean function $h_{[a_1,b_1],[a_2,b_2],[a_3,b_3]}$, where $x$ is positive for $h_{[a_1,b_1],[a_2,b_2],[a_3,b_3]}$ if $a_1 \leq x \leq b_1$ or $a_2 \leq x \leq b_2$ or $a_3 \leq x \leq b_3$ and $x$ is negative otherwise. Assume the intervals are disjoint. Give a polynomial-time PAC learning algorithm for learning the class $C_{3\text{INT}}$ of unions of 3 intervals in the realizable case. Also give an explicit sample complexity bound (you may use $O()$ notation).

2. VC-dimension and Rademacher Complexity.

   (a) Explain the importance of VC-dimension in machine learning.

   (b) Explain why the VC-dimension of any finite class $C$ is never greater than $\log_2 |C|$.

   (c) Give an example of an infinite concept class $C$ for which Sauer’s lemma is tight. That is, $C[m] = \sum_{i=0}^{d} \binom{m}{i}$ where $d$ is the VC-dimension of the class.

   (d) Explain when and why generalization bounds based on the Rademacher complexity can be tighter and better than those based on VC-dimension.

3. VC-dimension of specific classes. Consider the problem of learning the class of axis-parallel boxes with the origin as a corner. Specifically, let the instance space $X = \mathbb{R}^n$, and let $\text{Box}_n$ denote the class of axis-parallel boxes bounded between the origin and some point $a = (a_1, \ldots, a_n)$ in the positive orthant. That is, a target function $c_a$ is specified by a point $a \in \mathbb{R}_+^n$, and an example $x$ is positive iff $0 \leq x_i \leq a_i$ for all $i$.

   (a) What is the VC-dimension of this class? Argue both upper and lower bounds.

   (b) Give a number of examples that is sufficient to ensure that with probability $\geq 1 - \delta$, all $h \in \text{Box}_n$ satisfy $|\text{err}_D(h) - \text{err}_S(h)| \leq \epsilon$. You may use $O()$ notation.
4. **Online learning.** In lecture we saw that for the setting of prediction with expert advice, the expected number of mistakes $M$ of the Randomized Weighted Majority (RWM) algorithm satisfies:

$$M \leq \min_i \left[ -m_i \ln(1-\epsilon) + \ln n \right]$$

where $m_i$ is the number of mistakes of expert $i$ and $n$ is the total number of experts. Now, suppose that we have some prior belief $p$ over the experts about which we think is likely to be best. Show that if we initialize the weight of each expert $i$ to $p_i$ (rather than to 1) and then run RWM, the expected number of mistakes $M$ will satisfy:

$$M \leq \min_i \left[ -m_i \ln(1-\epsilon) + \ln(1/p_i) \right]$$

5. **Active learning.**

(a) Let $\mathcal{C}_{\text{circ}}$ be the class of origin-centered circles in $\mathbb{R}^2$. That is, $\mathcal{C}_{\text{circ}} = \{h_r : r \geq 0\}$ where we define $h_r(x) = 1$ if $||x|| \leq r$ and $h_r(x) = -1$ if $||x|| > r$. Show that using active learning, $\mathcal{C}_{\text{circ}}$ can be learned to error $\epsilon$ with probability $\geq 1-\delta$ from polynomially many unlabeled examples and just $O(\log 1/\epsilon)$ label requests. Hint: think about thresholds.

(b) Now, let $D$ be the uniform distribution over $\{x \in \mathbb{R}^2 : ||x|| = 1\}$, i.e., the unit circle in $\mathbb{R}^2$. Let $\mathcal{C}_{\text{ltf}}$ be the class of linear separators (not necessarily going through the origin). Show an $\Omega(1/\epsilon)$ lower bound on the number of label requests needed for active learning of $\mathcal{C}_{\text{ltf}}$ with respect to this distribution $D$. Hint: think about intervals.

6. **Equivalence queries.** In the equivalence query model of learning, we are given a concept class $\mathcal{C}$ and the goal of the learning algorithm is to exactly recover$^1$ the target function $c^*$. At each step, the learning algorithm can propose a hypothesis $h$ (which need not belong to $\mathcal{C}$) and then is given an example $x$ such that $h(x) \neq c^*(x)$ if such $x$ exists.

(a) Let $\mathcal{C}_k$ be the class of Boolean functions over $\{0, 1\}^n$ that have at most $k$ positive examples. Show how this class can be learned in the equivalence query model using at most $k$ equivalence queries.

(b) Consider the class of monotone conjunctions over $\{0, 1\}^n$. Show how this class can be learned in the equivalence query model using at most $n$ equivalence queries.

(c) Consider the class of decision lists over $\{0, 1\}^n$. Show how this class can be learned in the equivalence query model using $O(n^2)$ equivalence queries.

---

$^1$“Exactly recover” means to produce a function $h$ such that for all $x$ in the domain we have $h(x) = c^*(x)$. It does not require the functions to look syntactically the same.