COSC 480 Learning From Data
Midterm Exam
Fall 2019

Please print this document, write your *final* answers here, and upload to Gradescope. (You may want to draft answers on a separate document and copy them here.)

Collaboration is **not** permitted. You may use class notes, handouts, and the textbook for reference *and* for definitions and results that you may want to assume/use in your answers. However, you may **not** discuss the exam with classmates or use any other outside sources of information. You may of course ask the professor clarifying questions (please email me directly or come to office hours and I will share any common points of clarification to the class).

Please sign below to indicate that in completing this exam, you abided by the Colgate Academic Honor Code.

Name: _____

Signature: _____

Note Binomial coefficients will be useful in some of the questions, at least one maybe more (hint). Here are some useful identities that you may want to use:

- $\binom{N}{i} = \binom{N}{N-i}$
- $\sum_{i=0}^{N} \binom{N}{i} = 2^N$
- $\sum_{i=0}^{M} \binom{N}{i} = \sum_{i=0}^{M} \binom{N}{N-i} = \sum_{j=N-M}^{N} \binom{N}{j}$

1. In class, we looked at the Markov inequality, which states for any non-negative random variable X and any $\alpha > 0$,

$$Pr(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

Show that this bound is tight by describing a non-negative random variable X such that

$$Pr(X \geq \mathbb{E}[X]\,\alpha) = \frac{1}{\alpha}$$

2. Recall the concept $B(N, k)$, the maximum number of dichotomies on N points such that no subset of size k of N points can be shattered. Show that

$$B(N, k) \geq \sum_{i=0}^{k-1} \binom{N}{i}$$

(Hint: to show this lower bound you can construct a specific set of dichotomies. Think about the number of points labeled +1.)

3. Suppose you have two hypothesis sets $\mathcal{H}_A$ and $\mathcal{H}_B$ with VC dimension d_A and d_B respectively. You may assume that $d_A \leq d_B$. Let $\mathcal{H}_I \coloneqq \mathcal{H}_A \cap \mathcal{H}_B$ and $\mathcal{H}_U \coloneqq \mathcal{H}_A \cup \mathcal{H}_B$.

   (a) What is the *smallest* possible VC dimension for $\mathcal{H}_I$?

   (b) What is the *largest* possible VC dimension for $\mathcal{H}_I$?

   (c) What is the *smallest* possible VC dimension for $\mathcal{H}_U$?

(d) Show that the VC dimension of $\mathcal{H}_U$ is at most $d_A + d_B + 1$.

(e) Give a concrete example of $\mathcal{H}_A$ and $\mathcal{H}_B$ such that $\mathcal{H}_U$ achieves the upper bound from (d).

4. We are considering hypotheses of the form $h : \mathbb{R} \to \{-1, +1\}$. That is, classifiers that look at a single feature x. You are asked to analyze the VC dimension of various hypothesis sets. In each case, you can give an informal argument rather than a rigorous proof. But your answer must be complete (hint: showing the VC dimension is *at least* d is easier than showing the VC dimension is *exactly* d).

   (a) $h(x) = \text{sign}(w_1 x + w_0)$ where the weights are constrained $w_1 > 0$ and $w_0 < 0$.

   (b) $h(x) = \text{sign}(w_2 x^2 + w_1 x + w_0)$ where the weights are unconstrained.

5. Recall that the PLA works as follows. In each iteration, given the current weights $\mathbf{w}(t)$ it selects a misclassified point $(\mathbf{x}, y)$ and updates the weights as follows:

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) + y\mathbf{x}$$

PLA continues until no misclassified points remain.

This update does not account for how close $(\mathbf{x}, y)$ is to the current boundary. Intuitively, perhaps points that are close should cause a small update whereas points that are far should cause a larger update.

Here's a variant of the perceptron learning algorithm that takes closeness into account. Let $(\mathbf{x}(t), y(t))$ be the misclassified point selected on the t^{th} iteration and define $s(t) \coloneqq \mathbf{w}(t)\mathbf{x}(t)$.

The update rule is

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) + \eta(y - s(t))\mathbf{x}(t)$$

This algorithm continues until no misclassified points remain.

(a) Let $\mathbf{w}^*$ be a vector of weights that separates the data. Show that

$$\|\mathbf{w}(t) - \mathbf{w}^*\|^2 - \|\mathbf{w}(t+1) - \mathbf{w}^*\|^2 > 0$$

provided that $\eta < \frac{2(s^*(t) - s(t))}{(y - s(t))\|\mathbf{x}\|^2}$ where $s^*(t) \coloneqq \mathbf{w}^*\mathbf{x}(t)$.

In other words, on the t^{th} update, the weights move closer to $\mathbf{w}^*$ provided that the step size isn't too big.

space for answer continues...

(b) Define $M := \min_i y^{(i)}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$ and $R := \max_i \|\mathbf{x}^{(i)}\|$.

Argue that

$$\frac{2(s^*(t) - s(t))}{(y - s(t))\|\mathbf{x}\|^2} \geq \frac{M}{R^2}$$

Therefore if $\eta < M/R^2$, then this algorithm moves closer to $\mathbf{w}^*$ with each iteration.

6. After you complete the exam, please provide an estimate for how long you spent working on it.

6. _____

This page is intentionally blank. Label any work with the corresponding problem number.

This page is intentionally blank. Label any work with the corresponding problem number.