Homework 6

Spring 2021
(Due: Friday, Apr 16, 2021, 11:59 pm Eastern Time)

Please submit your homework through **gradescope**. You can write, scan, type, etc. But for the convenience of grading, please merge everything into a **single PDF**.

Objective

There are two things you will learn in this homework:

(a) Evaluate the VC dimension of some simple hypothesis sets.

(b) Derive the bias and variance of a simple linear model.

You will be asked some of these questions in Quiz 6. The Quiz will be open on Apr 17, 8am Eastern Time, and close on Apr 18, 8am Eastern Time. The Quiz is 30 minutes long.

Exercise 1. (VC dimension)
Compute the VC dimension of the following hypothesis sets.

(a) $\mathcal{H} = \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in [a,\infty), a \in \mathbb{R}\} \cup \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in (-\infty, a], a \in \mathbb{R}\}$. To clarify, the first subset is the positive ray, and the second subset is the negative ray. So the union is the set of all positive rays and negative rays.

(b) $\mathcal{H} = \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in [a,b], a, b \in \mathbb{R}\} \cup \{h : \mathbb{R} \to \{-1,+1\} | h(x) = -1, \forall x \in [a,b], a, b \in \mathbb{R}\}$. So this is the union of the positive intervals and the negative intervals.

(c) $\mathcal{H} = \{h : \mathbb{R}^2 \to \{-1,+1\} | h(\boldsymbol{x}) = +1, \forall \boldsymbol{x}$ where $\|\boldsymbol{x}\|_2 \leq b, \quad b \in \mathbb{R}\}$. Note that this is a *concentric* circle.

(d) $\mathcal{H} = \{h : \mathbb{R}^2 \to \{-1,+1\} | h(\boldsymbol{x}) = +1, \forall \boldsymbol{x}$ where $\|\boldsymbol{x} - \boldsymbol{a}\|_2 \leq b, \quad \boldsymbol{a} \in \mathbb{R}^2, b \in \mathbb{R}\}$. Note that this is a circle with an arbitrary center $\boldsymbol{a}$.

Exercise 2. (VC dimension)
Consider the following hypothesis set that has only one parameter $\alpha \in \mathbb{R}$:

$$\mathcal{H} = \{h_\alpha : \mathbb{R} \to \mathbb{R} | h_\alpha(x) = (-1)^{\lfloor \alpha x \rfloor}, \alpha \in \mathbb{R}\} \tag{1}$$

Prove that this hypothesis set has an infinite VC-dimension. Here, $\lfloor y \rfloor$ is the flooring operator which returns the closest integer smaller than or equal to y.

Hint: Recall that VC dimension requires you to know the growth function. The growth function is the worst case estimate of the number of dichotomies that can ever be created. So you need to construct a dataset containing $x_1, \ldots, x_N$ first. Move around these data points until you find the maximum number of dichotomies. The hint here is to consider $(x_1, x_2, ..., x_N) = (10^0, 10^1, ..., 10^{N-1})$. Say α is some number with at least $N-1$ decimal places. What do you notice about $\lfloor \alpha x_i \rfloor = \lfloor \alpha \times 10^i \rfloor$ for each i?

Exercise 3. (Bias and variance)
Consider a linear model such that

$$y_n = \boldsymbol{x}_n^T \boldsymbol{\theta} + e_n, \qquad n = 1, \ldots, N, \tag{2}$$

where $e_n \sim \text{Gaussian}(0, \sigma^2)$, or equivalently in the matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}. \tag{3}$$

Define the training dataset as $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$.

(a) Suppose we train the model by running a linear regression with the L_2-loss to obtain the predictor $g^{(\mathcal{D})}(\boldsymbol{x}') = \widehat{\boldsymbol{\theta}}^T \boldsymbol{x}'$ for some testing sample $\boldsymbol{x}'$. Express $g^{(\mathcal{D})}(\boldsymbol{x}')$ in terms of the $\boldsymbol{X}$, the testing sample $\boldsymbol{x}'$, the true model $\boldsymbol{\theta}$, and the error $\boldsymbol{e}$.

(b) Find the average predictor $\overline{g}(\boldsymbol{x}') = \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x}')]$. Is $g^{(\mathcal{D})}(\boldsymbol{x}') = \widehat{\boldsymbol{\theta}}^T \boldsymbol{x}'$ an unbiased estimator? Why?

(c) Derive the variance of the predictor $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\boldsymbol{x}') - \overline{g}(\boldsymbol{x}')\right)^2\right]$. Express your answer in terms of $\boldsymbol{X}$, $\boldsymbol{x}'$ and σ^2.

Exercise 4. (Project checkpoint # 6)
This is the final checkpoint for your project. As you can see, the amount of homework is reduced because I understand that you need more time to focus on your project. At this checkpoint I like to clarify the grading criteria.

- Every report will be reviewed by 3 TAs. You can treat them as the conference reviewers.
- Each TA will be asked to fill out a score sheet (30 points per TA):
  - (5 points) Completeness of the project. (5 = ready for a conference submission, 1 = bare minimum for this course.)
  - (5 points) How much understanding of the literature does the report show? (5 = you show an excellent understanding of the literature, 1 = you really don't know anything about the problem.)
  - (5 points) How significant is the proposed hypothesis? Basically, is it something important to the community, or has this been done already? (5 = important problem, 1 = trivial.)
  - (5 points) Does the report show any quantitative results to justify the proposed hypothesis? (5 = excellent quantitative results, 1 = no quantitative result at all.)
  - (5 points) Is the argument (theoretical and experiment) used to prove the proposed hypothesis valid? (5 = absolutely correct, 1 = many statements are flawed.)
  - (5 points) Quality of the writing. (5 = good writing at ICML level, 1 = unreadable.)
- The head TA will then decide the remaining 10 points based on the novelty of the project. Novel = very special, new, and innovative.
- Your total will be the sum of these scores.

The final project report is due on April 30, 2021, 11:59pm Eastern Time. Please submit through gradescope. All reports must be typed using the ICML template (in LaTeX). The page limit is 10 pages. References do not count towards the page limit.

You are welcome to collaborate with another student. However, you must acknowledge the other person in your report. Regardless whether you have collaborated with another person, you must write your own report.