Homework 6

Spring 2024
(Due: Friday, Apr 11, 2024, 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) Understand the concept of hypothesis set and why learning can be infeasible.

(b) Understand the limitation of Hoeffding inequality.

You will be asked some of these questions in Quiz 6.

Exercise 1.
Suppose that we have a learning scenario with 8 possible input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8$, each being a 3-bit binary vector. We are given a training dataset $\mathcal{D}$ that contains $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_5, y_5)\}$. Each label y_n is either $\circ$ or $\bullet$. The relationship between $\boldsymbol{x}_n$ and y_n is given by an unknown target function $f : \mathcal{X} \to \mathcal{Y}$. Since there are only three variables to be learned from data, there is a total of 2^3 possible f's we can possibly have. They are summarized in the figure below.

$\boldsymbol{x}_n$			y_n	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0	0	0	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
0	0	1	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
0	1	0	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
0	1	1	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
1	0	0	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
1	0	1		?	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
1	1	0		?	$\circ$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\circ$	$\bullet$	$\bullet$
1	1	1		?	$\circ$	$\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$

The following exercises involve different choices of the hypothesis set $\mathcal{H}$. You need to (i) Identify the final hypothesis g by listing the 8 entries it has for the 8 input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8$. For example, you can write $g = [\circ, \bullet, \bullet, \circ, \bullet, \circ, \bullet, \bullet]$. (ii) Compute how many of the 8 possible target functions agree with g on all the three out-sample points, on two of them, one one of them, and on none of them. For example, if $g = [\circ, \bullet, \bullet, \circ, \bullet, \circ, \bullet, \bullet]$, then it will match with three out-samples once (f_4), match with two out-samples three times (f_2, f_3, f_8), etc.

(a) $\mathcal{H}$ has only two hypotheses h_1 and h_2. The first hypothesis h_1 always return $\bullet$, and the second hypothesis h_2 always return $\circ$. The learning algorithm picks the hypothesis that matches the training set $\mathcal{D}$ the most.

(b) Same as (a), but the learning algorithm picks the hypothesis that matches the training set $\mathcal{D}$ the least.

(c) $\mathcal{H} = \{h\}$, where h is the XOR operation. That is, $h(\boldsymbol{x}) = \bullet$ if $\boldsymbol{x}$ contains an odd number of 1's and $h(\boldsymbol{x}) = \circ$ if $\boldsymbol{x}$ contains an even number of 1's.

Exercise 2.
In this exercise, we shall illustrate, with a simple numerical example, that given a hypothesis set $\mathcal{H} = \{h : \mathcal{X} \to \{+1, -1\}\}$ and samples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, if one does not let the hypothesis function $h \in \mathcal{H}$ be independent of the samples when computing the in-sample error E_{in}, then the probability $\mathbb{P}(|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon)$ does not necessarily obey Hoeffding's inequality. More specifically, for the final hypothesis g picked by the learning algorithm based on the training samples, $\mathbb{P}(|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon)$ does not necessarily satisfy the Hoeffding's inequality, and we indeed require the uniform bound.

Consider the following random experiment. Suppose we have 1000 fair coins. We flip each coin independently for $N = 10$ times. Let's focus on 3 coins as follows:

- coin_1 = the first coin flipped.
- $\text{coin}_{\text{rand}}$ = a coin you choose at random from the 1000 coins.
- coin_{min} = the coin that had the minimum frequency of heads. (You have 1000 coins and each is flipped 10 times. So one of the 1000 coins will have the minimum frequency of heads. In case of a tie, pick the earlier one).

Let V_1, V_{rand} and V_{min} be the fraction of heads we obtain for coin_1, $\text{coin}_{\text{rand}}$ and coin_{min} respectively.

(a) What is the probability of getting a head for coin_1, of getting a head for $\text{coin}_{\text{rand}}$ and of getting a head for coin_{min}? Denote them by μ_1, μ_{rand} and μ_{min}, respectively.

(b) In Python, repeat this entire experiment for 100,000 runs to get 100,000 instances of V_1, V_{rand} and V_{min}. Plot the histograms of the distributions of these three random variables.

(c) Using (b), plot the estimated $\mathbb{P}(|V_1 - \mu_1| > \epsilon)$, $\mathbb{P}(|V_{\text{rand}} - \mu_{\text{rand}}| > \epsilon)$ and $\mathbb{P}(|V_{\text{min}} - \mu_{\text{min}}| > \epsilon)$, together with the Hoeffding's bound $2\exp(-2\epsilon^2 N)$, for $\epsilon = 0, 0.05, 0.1, ..., 0.5$.

(d) Which coins obey the Hoeffding's bound, and which ones do not? Explain why.

Hint: Note that μ_1, μ_{rand} and μ_{min} are not necessarily equal to $\mathbb{E}[V_1]$, $\mathbb{E}[V_{\text{rand}}]$ and $\mathbb{E}[V_{\text{min}}]$ respectively. Pay particular attention to V_{min} and its $\mathbb{E}[V_{\text{min}}]$ and μ_{min}!

Exercise 3. (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 $\|\boldsymbol{x}\|_2 \leq b, \ \ b \in \mathbb{R}\}$. 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 $\|\boldsymbol{x} - \boldsymbol{a}\|_2 \leq b, \ \ \boldsymbol{a} \in \mathbb{R}^2, b \in \mathbb{R}\}$. Note that this is a circle with an arbitrary center $\boldsymbol{a}$.

Exercise 4. (BIAS AND VARIANCE)
Consider a linear model such that

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

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}. \tag{2}$$

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 5. Project Check Point 4
On Gradescope, there is an assignment called Project Check Point 4. Submit a one-page manuscript that describe what real world problem that you propose to solve using your reimplementation, where the data comes from, and what the intended outcome is. Due at Homework 5 deadline.

The format is the same as the Check Point 1. Each team only needs to upload one copy, but make sure to list all group members of the team. The manuscript needs to be written using Latex following the ICML template listed on the course website.

The Check Point 4 will be graded separately from your HW5 submission, and is worth 15% of the final project grade. The grading will be based on the following criteria:

- Did the team describe clearly the problem they aim to solve with all necessary details well organized?
- Is the problem meaningful and related to the method in the paper?
- Is the report well written? Did the team copy or closely resemble the original paper or other people's reimplementation? Heavy penalty will be added to the manuscripts that copies or closely resembles the original paper or other people's reimplementation.