Homework 5

Spring 2021
(Due: Friday, Apr 2, 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 three 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.

(c) Practice on how to derive a Chernoff bound.

You will be asked some of these questions in Quiz 5. The Quiz will be open on Apr 3, 8am Eastern Time, and close on Apr 4, 8am Eastern Time. The Quiz is 30 minutes long.

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.

(d) $\mathcal{H}$ = all Boolean functions of three variables. The learning algorithm picks the hypothesis that agree with the training set, but disagree with the XOR.

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: The law of total probability could be useful here.

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.
Let $X_1, \ldots, X_N$ be i.i.d. Bernoulli random variables with $X_n \sim \text{Bernoulli}(0.5)$. Let $\overline{X}_N = (1/N)\sum_{n=1}^N X_n$ be the sample average.

(a) Use Chernoff bound to show that

$$\mathbb{P}[\overline{X}_N - \mu \geq \epsilon] \leq 2^{-\beta N}, \tag{1}$$

where $\beta = 1 + (\frac{1}{2} + \epsilon)\log_2(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon)\log_2(\frac{1}{2} - \epsilon)$, and $\mu = 0.5$.

(b) Write a Python program to verify the tightness of this bound with the Hoeffding inequality we demonstrated in class. [1] You need to modify the lines `prob_simulate[i]` and `prob_hoeffding[i]` to match with this problem. Remember to add legends for each curve. So altogether, you will have a set of crosses for the simulated points. A curve for Hoeffiding, and a curve for Chernoff.

[1] `https://engineering.purdue.edu/ChanGroup/ECE595/files/ECE595_demo_15.html`

Exercise 4. Project Checkpoint #5

Phase 1 submission (optional). If you read the project instructions, you will notice that we have an optional phase 1 submission.

`https://engineering.purdue.edu/ChanGroup/ECE595/project.html`

You can submit your tentative progress report by Apr 9, 2021, 11:59pm Eastern Time through gradescope. The TAs will read your report and try to offer some constructive feedback. The difference between this submission and the usual homework checkpoints is that in the homework checkpoints, we only check for completeness. Our goal is to ensure that you are on track. If you are lagging, we will drop you a note. For this phase 1 submission, if you choose to submit, the TAs will read and provide technical feedback. You may treat this as an opportunity to make your project better. To submit to phase 1, we expect that your report contains some meat. It does not need to be complete, but it must have something.

Check point # 5. At this checkpoint, I assume that you have already worked on your problem for some time. There are three possibilities I can foresee: (1) You are struggling with the code or equations. This is not a problem. Please continue. If you need help, feel free to reach out to our TAs. (2) You have been able to implement something, and you have formulated your hypothesis. You are working on the experiments. This is excellent, please continue. Please consider phase 1 submission. (3) You switched from one topic to another topic, and you still haven't settled down on what you would want to do. Or you have settled down on a topic, but you have not yet started working on things. This is not good. Please catch up.

No matter which situation you are in, I have a few general pieces of advice:

- It is never too early to start thinking about your project report. I have some expectations of what you need to report. Specifically, I want to see: (1) A clear statement of your hypothesis, and justifications of why this is an interesting problem. (2) A thorough discussion of the literature, including things people have done and their limitations. (3) A well-developed set of arguments and experiments to justify your hypothesis. For details, please check the project grading criteria stated on the course website.

- Always reports **quantitative experiments**. Explain how the experiments were set up. Explain the configurations. Explain the significance of the experiment. Explain why this experiment can justify the hypothesis you are claiming.

- If you are working on a hypothesis that involves comparing with other methods, remember to explain why that method is a reasonable baseline. It is okay that you are not able to beat the baseline. But you need to report them. Explain why you are not able to beat them. Suggest possible solutions.

- As I always say, please focus on one hypothesis. The hypothesis can be simple. You do not need to be too ambitious. As long as you can justify why this hypothesis is important, we will accept it. However, the hypothesis cannot be trivial. This includes things that have already been proven in the literature and there are zero controversies about the statement. For example, noise2noise performance drops when noise becomes stronger.

- I strongly encourage you to work within the scope of the suggested topics. While I have the flexibility to things slightly outside the bounding box of the four topics, they cannot be completely off the track. Each topic has its key spirit. E.g., Noise2noise is unsupervised learning. No matter what you do, it must be about unsupervised learning. PnP is PnP. No matter what you do, it must be within the framework of PnP. An attack is an attack. You can play around with the proportion of the samples, but it has to be about an attack. A noisy label is a noisy label. No matter what you do, it has to tackle noisy labels.

