ECE 50024 / STAT 59800: Machine Learning I
Spring 2023
Instructor: Prof. Qi Guo, Developer: Prof. Stanley H Chan

Homework 6

Spring 2023
(Due: Friday, Apr 20, 2023, 4:59 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.

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	○	○	○	○	○	○	○	○	○	○
0	0	1	●	●	●	●	●	●	●	●	●	●
0	1	0	●	●	●	●	●	●	●	●	●	●
0	1	1	○	○	○	○	○	○	○	○	○	○
1	0	0	●	●	●	●	●	●	●	●	●	●
1	0	1		?	○	○	○	○	●	●	●	●
1	1	0		?	○	○	●	●	○	○	●	●
1	1	1		?	○	●	○	●	○	●	○	●

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. 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.

The final project report is due on Apr 30, 2022, 4:59pm 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.

Also, please upload your code to a github repository and put the link to the repository in your report. If you want to keep your repository private, you can upload to Purdue's internal Github at `https://github.itap.purdue.edu/` as a private repository and add the teaching staff as collaborators.

- Every report will be reviewed by 3 teaching staff. You can treat them as the conference reviewers.
- Each teaching staff will be asked to fill out a score sheet based on the following criteria listed on the course website. Your score will be the average of the scores from three independent reviewers.
  - Does the paper clearly state the problem that the implemented machine learning model targets? You will have to use your own language to describe the problem. Heavy penalty will be added for copying (with moderate modification of) the original paper. Does the paper identify and clearly descirbe similar works in the related work, and lists their advantages and disadvantages?
  - Does the paper explain the mathematical derivation well? You will have to use your own language to form the mathematical derviation. Heavy penalty will be added for copying (with moderate modification of) the original paper.

- Does the paper clearly state the technical difficulties in the reimplementation and possible solutions?
- Does the paper identify, describe, and analyze the effect of different engineering practices, e.g., implementation tricks and parameter/hyperparameter choices, using scientific languages?
- Does the student train the implemented model using the dataset they select, and identify and solve issues that prevent the model from convergence?
- Does the paper analyze the experimental behaviors of their model on selected application and compare with baselines?
- Are there sufficient experiments demonstrating the success of re-implementation?
- Are the limitations/assumptions clearly stated in the paper? Are there overclaim? Heavy penalty will be added for overclaim.
- The amount of work in the reimplementation.
- Is the paper easy to read? Are your ideas elaborated clearly?