Homework #2

RELEASE DATE: 09/27/2023

DUE DATE: 10/18/2023, BEFORE 13:00 on GRADESCOPE

QUESTIONS ARE WELCOMED ON DISCORD (INFORMALLY) OR NTU COOL (FORMALLY).

You will use Gradescope to upload your scanned/printed solutions. For problems marked with (*), please follow the guidelines on the course website and upload your source code to Gradescope as well. Any programming language/platform is allowed.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.

Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.

You should write your solutions in English or Chinese with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

This homework set comes with 240 points and 20 bonus points. In general, every homework set would come with a full credit of 240 points, with some possible bonus points.

Hoeffding Inequalities

1. (20 points) Assume that we have M slot machines in front of us. Each machine has an unknown probability of μ_m for returning one coin, and a probability of $1 - \mu_m$ for returning no coin. For each of the time step $t = 1, 2, \ldots$, assume that we pull the machine $m = ((t-1) \bmod M) + 1$. After some $t > M$ time steps, we'd have pulled machine m for N_m times, and collected c_m coins from machine m. Note that $N_m \geq 1$ because $t > M$. Consider the following one-sided Hoeffding's inequality (which is slightly different from what we taught in class)

$$P\left(\mu > \nu + \epsilon\right) \leq \exp(-2\epsilon^2 N),$$

where ν, μ, ϵ, N have been defined in our class. Use the inequality above to prove that when given a fixed machine m and a fixed δ with $0 < \delta < 1$,

$$P\left(\mu_m > \frac{c_m}{N_m} + \sqrt{\frac{\log t - \frac{1}{2}\log\delta}{N_m}}\right) \leq \delta t^{-2}.$$

2. (20 points) Continuing from Problem 1, prove that when $M \geq 2$, for all slot machines $m = 1, 2, \ldots, M$ and for all $t = M+1, M+2, \ldots$, with probability at least $1 - \delta$,

$$\mu_m \leq \frac{c_m}{N_m} + \sqrt{\frac{\log t + \log M - \frac{1}{2}\log\delta}{N_m}}.$$

You can use the magical fact that

$$\sum_{t=1}^{\infty} t^{-2} = \frac{\pi^2}{6}.$$

Hint: The fact that we can upper-bound all μ_m confidently and simultaneously by $\frac{c_m}{N_m}$ plus a deviation term is the core technique for deriving the so-called upper-confidence bound algorithm for

multi-armed bandits, which is an important algorithm for the task of online and reinforcement learning. The actual algorithm differs from what we do in Problem 1 by pulling the machine with the largest upper confidence bound in each iteration, instead of periodically going through each machine. Those who are interested can certainly search for more about this.

3. (20 points) Next, we illustrate what happens with multiple bins. Consider a special lottery game as follows. The game operates by having four kinds of lottery tickets placed in a big black bag, each kind with the same (super large) quantity. Exactly eight numbers $1, 2, \ldots, 8$ are written on each ticket. The four kinds are

- A: all even numbers are colored orange, all odd numbers are colored green
- B: all even numbers are colored green, all odd numbers are colored orange
- C: all small numbers (1-4) are colored orange, all big numbers (5-8) are colored green
- D: all small numbers (1-4) are colored green, all big numbers (5-8) are colored orange

Every person is expected to draw five tickets from the bag. A small price of 1450 is given if the five tickets contain "some number" that is purely green. What is the probability that such an event will happen?

4. (20 points) Continuing from Problem 3, a bigger price of three piggy banks will be given if the five tickets contain five green 2's. What is the probability that such an event will happen?

Hint: Each number can be viewed as a "hypothesis" and the drawn tickets can be viewed as the data. The E_{out} of each hypothesis is simply $\frac{1}{2}$ (You are welcome. ;-)). Problem 4 asks you to calculate the BAD probability for hypothesis 2; Problem 3 asks you to calculate the BAD probability for all hypotheses, taking the dependence into consideration.

VC Dimension

5. (20 points) Consider the "negative rectangle" hypothesis set for $\mathcal{X} = \mathbb{R}^2$, which includes any hypothesis that returns -1 when $\mathbf{x}$ is within an axis-parallel rectangle and $+1$ elsewhere. Show that some set of 4 input vectors can be shattered by the hypothesis set. That is, the VC dimension of the hypothesis set is no less than 4.

6. (20 points) Consider a hypothesis set $\mathcal{H}$ for $\mathcal{X} = \mathbb{R}$ containing hypothesis with $2M + 1$ ($M \geq 1$) parameters. Each hypothesis $h(x)$ in $\mathcal{H}$ are defined by $s, a_1, b_1, a_2, b_2, \ldots, a_M, b_M$ that satisfies

- $s \in \{+1, -1\}$
- $a_m < b_m$, for $1 \leq m \leq M$;
- $b_m < a_{m+1}$, for $1 \leq m \leq M - 1$,

with

$$h_{s,\mathbf{a},\mathbf{b}}(x) = \begin{cases} s, \text{ if } a_m \leq x \leq b_m \text{ for some } 1 \leq m \leq M \\ -s, \text{ otherwise} \end{cases}$$

What is the VC dimension of $\mathcal{H}$? Prove your answer.

Hint: The positive intervals introduced in Lecture 5 correspond to $s = +1$ with $M = 1$.

7. (20 points) What is the growth function of origin-passing perceptrons on $\mathcal{X} = \mathbb{R}^2$? Those perceptrons are

$$\mathcal{H}_0 \quad = \quad \{h \colon h(\mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2) \quad \textit{i.e.} \text{ perceptrons that pass the origin}\}$$

Prove your answer.

Hint: Consider putting your input vectors on the unit circle.

8. (20 points) For $\mathcal{X} = \mathbb{R}^2$, consider a hypothesis set $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$ that is a union of two types of perceptrons:

$$\mathcal{H}_0 \quad = \quad \{h \colon h(\mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2) \quad \textit{i.e.} \text{ perceptrons that pass the origin}\}$$
$$\mathcal{H}_1 \quad = \quad \{h \colon h(\mathbf{x}) = \text{sign}(w_1 (x_1 - 1) + w_2 (x_2 - 1)) \quad \textit{i.e.} \text{ perceptrons that pass } (1, 1)\}$$

What is the VC dimension of $\mathcal{H}$? Prove your answer.

Decision Stumps

9. (20 points) In class, we taught about the learning model of "positive and negative rays" (which is simply one-dimensional perceptron) for one-dimensional data. The model contains hypotheses of the form:

$$h_{s,\theta}(x) = s \cdot \text{sign}(x - \theta).$$

You can take $\text{sign}(0) = -1$ for simplicity but it should not matter much for the following problems. The model is frequently named the "decision stump" model and is one of the simplest learning models. As shown in class, for one-dimensional data, the VC dimension of the decision stump model is 2.

In the following problems, you are asked to play with decision stumps on an artificial data set. First, start by generating a one-dimensional data by the procedure below:

a) Generate x by a uniform distribution in $[-1, 1]$.

b) Generate y by $y = \text{sign}(x) + \text{noise}$, where the noise flips the sign with 10% probability.

With the (x, y) generation process above, prove that for any $h_{s,\theta}$ with $s \in \{-1, +1\}$ and $\theta \in [-1, 1]$,

$$E_{\text{out}}(h_{s,\theta}) = 0.5 - 0.4s + 0.4s \cdot |\theta|.$$

10. (20 points, *) In fact, the decision stump model is one of the few models that we could minimize E_{in} efficiently by enumerating all possible thresholds. In particular, for N examples, there are at most $2N$ dichotomies (see the slides for positive rays), and thus at most $2N$ different E_{in} values. We can then easily choose the hypothesis that leads to the lowest E_{in} by the following decision stump learning algorithm.

(1) sort all N examples x_n to a sorted sequence $x'_1, x'_2, \ldots, x'_N$ such that $x'_1 \le x'_2 \le x'_3 \le \ldots \le x'_N$

(2) for each $\theta \in \{-1\} \cup \{\frac{x'_i + x'_{i+1}}{2} : 1 \le i \le N-1 \text{ and } x'_i \ne x'_{i+1}\}$ and $s \in \{-1, +1\}$, calculate $E_{\text{in}}(h_{s,\theta})$

(3) return the $h_{s,\theta}$ with the minimum E_{in} as g; if multiple hypotheses reach the minimum E_{in}, return the one with the smallest $s \cdot \theta$.

(*Hint: CS-majored students are encouraged to think about whether the second step can be carried out efficiently, i.e. $O(N)$, using dxxxxc pxxxxxxxxg instead of the naive implementation of $O(N^2)$.*)

Generate a data set of size 32 by the procedure above and run the one-dimensional decision stump algorithm on the data set to get g. Record $E_{\text{in}}(g)$ and compute $E_{\text{out}}(g)$ with the formula in Problem 9. Repeat the experiment 2000 times. Plot a scatter plot of $(E_{\text{in}}(g), E_{\text{out}}(g))$, and calculate the median of $E_{\text{out}}(g) - E_{\text{in}}(g)$.

11. (20 points, *) Repeat Problem 10, but generate a data set of size 8 by the procedure instead. Plot a scatter plot of $(E_{\text{in}}(g), E_{\text{out}}(g))$, and calculate the median of $E_{\text{out}}(g) - E_{\text{in}}(g)$. Compare the scatter plot and the median value with those of Problem 10. Describe your findings.

12. (20 points, *) Repeat Problem 11, generate a data set of size 8 by the procedure above. Instead of running the decision stump algorithm, return a randomly chosen $h_{s,\theta}$ as g, with s uniformly sampled from $\{-1, +1\}$ and θ uniformly sampled from $[-1, 1]$. Record $E_{\text{in}}(g)$ and compute $E_{\text{out}}(g)$ with formula in Problem 9. Repeat the experiment 2000 times. Plot a scatter plot of $(E_{\text{in}}(g), E_{\text{out}}(g))$, and calculate the median of $E_{\text{out}}(g) - E_{\text{in}}(g))$. Compare the scatter plot and the median value with those of Problem 11. Describe your findings.

Bonus: Perceptrons that Pass Special Points

13. (Bonus 20 points) Consider $\mathcal{H}$ being perceptrons in $\mathcal{X} = \mathbb{R}^d$. It is known, by the so-called Cover's Theorem, that the growth function is

$$m_{\mathcal{H}}(N) = 2\sum_{i=0}^{d} \binom{N-1}{i}.$$

See, for instance,

`https://web.mit.edu/course/other/i2course/www/vision_and_learning/perceptron_notes.pdf`

for its proof.

Now, assume that we require the perceptrons to pass *all* k anchor points for $\mathbf{a}_1$, $\mathbf{a}_2$, …, $\mathbf{a}_k$, each being in $\mathbb{R}^d$ with $0 \leq k < d$. We shall call those perceptrons $\tilde{\mathcal{H}}$. What is the growth function $m_{\tilde{\mathcal{H}}}(N)$? Prove your answer.

Note: Problem 7 is a special case for $k = 1$ *and* $\mathbf{a}_1 = \mathbf{0}$.