**Homework #2**

RELEASE DATE: 10/16/2020

RED BUG FIX: 10/22/2020 16:30

BLUE BUG FIX: 10/24/2020 17:00

DUE DATE: EXTENDED TO 11/06/2020, BEFORE 13:00 on NTU COOL

QUESTIONS ARE WELCOMED ON THE NTU COOL FORUM.

*We will instruct you on how to use Gradescope to upload your choices and your scanned/printed solutions later. For problems marked with (*), please follow the guidelines on the course website and upload your source code to Gradescope as well. You are encouraged to (but not required to) include a README to help the TAs check your source code. 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 400 points. For each problem, there is one correct choice. For most of the problems, if you choose the correct answer, you get 20 points; if you choose an incorrect answer, you get $-10$ points. That is, the expected value of random guessing is $-20$ per problem, and if you can eliminate two of the choices accurately, the expected value of random guessing on the remaining three choices would be 0 per problem. For other problems, the TAs will check your solution in terms of the written explanations and/or code. The solution will be given points between $[-20, 20]$ based on how logical your solution is.

## Perceptrons

**1.** Which of the following set of $\mathbf{x} \in \mathbb{R}^3$ can be shattered by the 3D perceptron hypothesis set? The set contains all hyperplanes of the form with our usual notation of $x_0 = 1$:

$$h_{\mathbf{w}}(\mathbf{x}) = \text{sign}\left(\sum_{i=0}^{3} w_i x_i\right).$$

Choose the correct answer; explain your answer.

**[a]** $\{(7,8,9),(17,18,19),(27,28,29)\}$

**[b]** $\{(1,1,1),(7,8,9),(15,16,17),(21,23,25)\}$

**[c]** $\{(1,1,3),(7,8,9),(15,16,17),(21,23,25)\}$

**[d]** $\{(1,3,5),(7,8,9),(15,16,17),(21,23,25)\}$

**[e]** $\{(1,2,3),(4,5,6),(7,8,9),(15,16,17),(21,23,25)\}$

**2.** What is the growth function of axis-aligned perceptrons in 2D for $N \geq 4$? Those perceptrons are all perceptrons with $w_1 w_2 = 0$. That is, they are vertical or horizontal lines on the 2D plane. Choose the correct answer; explain your answer.

**[a]** $4N + 4$

**[b]** $4N + 2$

**[c]** $4N$

**[d]** $4N - 2$

**[e]** $4N - 4$

**3.** What is the VC dimension of positively-biased perceptrons in 2D? The positively-biased perceptrons are all perceptrons with $w_0 > 0$. Choose the correct answer; explain your answer.

**[a]** 0

**[b]** 1

**[c]** 2

**[d]** 3

**[e]** 4

## Ring Hypothesis Set

**4.** The "ring" hypothesis set in $\mathbb{R}^3$ contains hypothesis parameterized by two positive numbers $a$ and $b$, where

$$h(\mathbf{x}) = \begin{cases} +1 & \text{if } a \leq x_1^2 + x_2^2 + x_3^2 \leq b, \\ -1 & \text{otherwise.} \end{cases}$$

What is the growth function of the hypothesis set? Choose the correct answer; explain your answer.

**[a]** $\binom{N+1}{1} + 1$

**[b]** $\binom{N+1}{2} + 1$

**[c]** $\binom{N+1}{3} + 1$

**[d]** $\binom{N+1}{6} + 1$

**[e]** none of the other choices

**5.** Following the previous problem, what is the VC dimension of the ring hypothesis set? Choose the correct answer; explain your answer.

**[a]** 1

**[b]** 2

**[c]** 3

**[d]** 6

**[e]** none of the other choices

## Deviation from Optimal Hypothesis

**6.** In Lecture 7, the VC bound was stated from the perspective of $g$, the hypothesis picked by the learning algorithm. The bound itself actually quantifies the BAD probability from *any* hypothesis $h$ in the hypothesis set. That is,

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 4 m_{\mathcal{H}}(2N) \exp\left(-\frac{1}{8}\epsilon^2 N\right).$$

Define the best-$E_{\text{in}}$ hypothesis

$$g = \text{argmin}_{h \in \mathcal{H}} E_{\text{in}}(h)$$

and the best-$E_{\text{out}}$ hypothesis (which is optimal but can only be obtained by a "cheating" algorithm)

$$g_* = \text{argmin}_{h \in \mathcal{H}} E_{\text{out}}(h).$$

Using the VC bound above, with probability more than $1 - \delta$, which of the following is an upper bound of $E_{\text{out}}(g) - E_{\text{out}}(g_*)$? Choose the correct answer; explain your answer.

**[a]** $\sqrt{\frac{1}{8N} \ln\left(\frac{4 m_{\mathcal{H}}(2N)}{\delta}\right)}$

**[b]** $\sqrt{\frac{1}{8N} \ln\left(\frac{m_{\mathcal{H}}(2N)}{4\delta}\right)}$

**[c]** $\sqrt{\frac{8}{N} \ln\left(\frac{4 m_{\mathcal{H}}(2N)}{\delta}\right)}$

**[d]** $2\sqrt{\frac{8}{N} \ln\left(\frac{4 m_{\mathcal{H}}(2N)}{\delta}\right)}$

**[e]** $\sqrt{\frac{8}{N} \ln\left(\frac{8 m_{\mathcal{H}}(2N)}{\delta}\right)}$

## The VC Dimension

**7.** For a finite hypothesis set $\mathcal{H} = \{h_1, h_2, \ldots, h_M\}$, where each hypothesis is a binary classifier from $\mathcal{X}$ to $\{-1, +1\}$, what is the largest possible value of $d_{\text{vc}}(\mathcal{H})$? Choose the correct answer; explain your answer.

**[a]** $M$

**[b]** $2M$

**[c]** $M^2$

**[d]** $\lfloor \log_2 M \rfloor$

**[e]** $2^M$

**8.** A boolean function $h\colon \{-1, +1\}^k \to \{-1, +1\}$ is called *symmetric* if its value does not depend on the permutation of its inputs, i.e., its value only depend on the number of ones in the input. What is the VC dimension of the set of all symmetric boolean functions? Choose the correct answer; explain your answer.

**[a]** $k - 2$

**[b]** $k - 1$

**[c]** $k$

**[d]** $k + 1$

**[e]** $k + 2$

**9.** How many of the following are necessary conditions for $d_{\text{vc}}(H) = d$? Choose the correct answer; state which conditions correspond your answer and explain them.

- some set of $d$ distinct inputs is shattered by $\mathcal{H}$
- some set of $d$ distinct inputs is not shattered by $\mathcal{H}$
- any set of $d$ distinct inputs is shattered by $\mathcal{H}$
- any set of $d$ distinct inputs is not shattered by $\mathcal{H}$
- some set of $d+1$ distinct inputs is shattered by $\mathcal{H}$
- some set of $d+1$ distinct inputs is not shattered by $\mathcal{H}$
- any set of $d+1$ distinct inputs is shattered by $\mathcal{H}$
- any set of $d+1$ distinct inputs is not shattered by $\mathcal{H}$

**[a]** 1

**[b]** 2

**[c]** 3

**[d]** 4

**[e]** 5

**10.** Which of the following hypothesis set is of VC dimension $\infty$? Choose the correct answer; explain your answer.

**[a]** the rectangle family: the infinite number of hypotheses where $h(\mathbf{x}) = 0$ looks like a rectangle (including axis-aligned ones and rotated ones) for $\mathbf{x} \in \mathbb{R}^2$

**[b]** the intersected-interval family: the infinite number of hypotheses where the positive region of each hypothesis can be represented as an intersection of any finite number of "positive intervals" for $x \in \mathbb{R}$

**[c]** the sine family: the infinite number of hypotheses $\{h_\alpha \colon h_\alpha(x) = \text{sign}(\sin(\alpha \cdot x))\}$ for $x \in \mathbb{R}$

**[d]** the scaling family: the infinite number of hypothesis $\{h_\alpha \colon h_\alpha(\mathbf{x}) = \text{sign}(\alpha \cdot \sum_{i=1}^{d} x_i)\}$ for $\mathbf{x} \in \mathbb{R}^d$

**[e]** none of the other choices

## Noise and Error

**11.** Consider a binary classification problem where we sample $(\mathbf{x}, y)$ from a distribution $\mathcal{P}$ with $y \in \{-1, +1\}$. Now we define a distribution $\mathcal{P}_\tau$ to be a "noisy" version of $\mathcal{P}$. That is, to sample from $\mathcal{P}_\tau$, we first sample $(\mathbf{x}, y)$ from $\mathcal{P}$ and flip $y$ to $-y$ with probability $\tau$ independently. Note that $\mathcal{P}_0 = \mathcal{P}$. The distribution $\mathcal{P}_\tau$ models a situation that our training data is labeled by an unreliable human, who mislabels with probability $\tau$.

Define $E_{\text{out}}(h, \tau)$ to be the out-of-sample error of $h$ with respect to $\mathcal{P}_\tau$. That is,

$$E_{\text{out}}(h, \tau) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_\tau} \left[\!\left[ h(\mathbf{x}) \neq y \right]\!\right].$$

Which of the following relates $E_{\text{out}}(h, \tau)$ to $E_{\text{out}}(h, 0)$? Choose the correct answer; explain your answer.

**[a]** $E_{\text{out}}(h, 0) = \frac{E_{\text{out}}(h,\tau) - 2\tau}{1-\tau}$

**[b]** $E_{\text{out}}(h, 0) = \frac{2E_{\text{out}}(h,\tau) - \tau}{2-\tau}$

**[c]** $E_{\text{out}}(h, 0) = \frac{\tau - E_{\text{out}}(h,\tau)}{1-2\tau}$

**[d]** $E_{\text{out}}(h, 0) = \frac{E_{\text{out}}(h,\tau) - \tau}{1-2\tau}$

**[e]** $E_{\text{out}}(h, 0) = \frac{2E_{\text{out}}(h,\tau) - 2\tau}{2-\tau}$

**12.** Consider $\mathbf{x} \in \mathbb{R}^3$ and a target function $f(\mathbf{x}) = \text{argmax}_{i=1,2,3} x_i$, with ties broken, if any, by choosing the smallest $i$. Then, assume a process that generates $(\mathbf{x}, y)$ by a uniform $P(\mathbf{x})$ within $[0,1]^3$ and

$$P(y|\mathbf{x}) = \begin{cases} 0.7 & y = f(\mathbf{x}) \\ 0.1 & y = f(\mathbf{x}) \bmod 3 + 1 \\ 0.2 & y = (f(\mathbf{x}) + 1) \bmod 3 + 1 \end{cases}$$

The operation of "$a$ mod 3" returns the residual when the integer $a$ is divided by 3. When using the squared error, what is $E_{\text{out}}(f)$ subject to the process above? Choose the correct answer; explain your answer. (*Note: This is in some sense the "price of noise"*)

**[a]** 0.3

**[b]** 0.6

**[c]** 0.9

**[d]** 1.2

**[e]** 1.5

**13.** Following Problem 12, the squared error defines an ideal target function

$$f_*(\mathbf{x}) = \sum_{y=1}^{3} y \cdot P(y|\mathbf{x}),$$

as shown on page 11 of the Lecture 8 slides. Unlike the slides, however, we denote this function as $f_*$ to avoid being confused with the target function $f$ used for generating the data. Define the squared difference between $f$ and $f_*$ to be

$$\Delta(f, f_*) = \mathbb{E}_{\mathbf{x} \sim P(\mathbf{x})} (f(\mathbf{x}) - f_*(\mathbf{x}))^2.$$

What is the value of $\Delta(f, f_*)$? Choose the correct answer; explain your answer. (*Note: This means how much the original target function $f$ was dragged by the noise in $P(y|\mathbf{x})$ to the "new" target function $f_*$.*)

**[a]** 0.01

**[b]** 0.14

**[c]** 0.16

**[d]** 0.25

**[e]** 0.42

## Decision Stump

In page 22 of the Lecture 5 slides (the Fun Time that you should play by yourself), we taught about the learning model of "positive and negative rays" (which is simply one-dimensional perceptron). The model contains hypotheses of the form:

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

where $s \in \{-1, +1\}$ is the "direction" of the ray and $\theta \in \mathbb{R}$ is the threshold. You can take $\text{sign}(0) = -1$ for simplicity. The model is frequently named the "decision stump" model and is one of the simplest learning models. As shown in class, the growth function of the model is $2N$ and the VC Dimension is 2.

**14.** When using the decision stump model, given $\epsilon = 0.1$ and $\delta = 0.1$, among the five choices, what is the smallest $N$ such that the BAD probability of the VC bound (as given in the beginning of Problem 6) is $\leq \delta$? Choose the correct answer; explain your answer.

**[a]** 6000
**[b]** 8000
**[c]** 10000
**[d]** 12000
**[e]** 14000

In fact, the decision stump model is one of the few models that we could minimize $E_{\text{in}}$ efficiently by enumerating all possible thresholds. In particular, for $N$ examples, there are at most $2N$ dichotomies (see page 22 of the Lecture 5 slides), and thus at most $2N$ different $E_{\text{in}}$ values. We can then easily choose the hypothesis that leads to the lowest $E_{\text{in}}$ 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 $x'_1 \leq x'_2 \leq x'_3 \leq \ldots \leq x'_N$

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

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

(*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 $O(N^2)$.*)

Next, you are asked to implement such an algorithm and run your program on an artificial data set. We shall start by generating $(x, y)$ with the following procedure. We will take the target function $f(x) = \text{sign}(x)$:

- Generate $x$ by a uniform distribution in $[-1, +1]$.
- Generate $y$ from $x$ by $y = f(x)$ and then flip $y$ to $-y$ with $\tau$ probability independently

**15.** For $\theta \in [-1, +1]$, what is $E_{\text{out}}(h_{+1,\theta}, 0)$, where $E_{\text{out}}(h, \tau)$ is defined in Problem 11? Choose the correct answer; explain your answer.

**[a]** $|\theta|$
**[b]** $\frac{1}{2}|\theta|$
**[c]** $2|\theta|$
**[d]** $1 - |\theta|$
**[e]** $1 - \frac{1}{2}|\theta|$

**16.** (*) For $\tau = 0$, which means that your data is noiseless. Generate a data set of size by the procedure above and run the decision stump algorithm on the data set to get $g$. Repeat the experiment 10000 times, each with a different data set. What is the mean of $E_{\text{out}}(g, \tau) - E_{\text{in}}(g)$ within the 10000 results? Choose the closest value. (*By extending the results in Problem 11 and Problem 15, you can actually compute any $E_{\text{out}}(h_{s,\theta}, \tau)$ analytically. But if you do not trust your math derivation, you can get a very accurate estimate of $E_{\text{out}}(g)$ by evaluating $g$ on a separate test data set of size* 100000*, as guaranteed by Hoeffding's inequality*).

**[a]** 0.00
**[b]** 0.02
**[c]** 0.05
**[d]** 0.30
**[e]** 0.40

**17.** (*) For $\tau = 0$, generate a data set of size 20 by the procedure above and run the decision stump algorithm on the data set to get $g$. Repeat the experiment 10000 times, each with a different data set. What is the mean of $E_{\text{out}}(g, \tau) - E_{\text{in}}(g)$ within the 10000 results? Choose the closest value.

**[a]** 0.00
**[b]** 0.02
**[c]** 0.05
**[d]** 0.30
**[e]** 0.40

**18.** (*) For $\tau = 0.1$, generate a data set of size 2 by the procedure above and run the decision stump algorithm on the data set to get $g$. Repeat the experiment 10000 times, each with a different data set. What is the mean of $E_{\text{out}}(g, \tau) - E_{\text{in}}(g)$ within the 10000 results?

**[a]** 0.00
**[b]** 0.02
**[c]** 0.05
**[d]** 0.30
**[e]** 0.40

**19.** (*) For $\tau = 0.1$, generate a data set of size 20 by the procedure above and run the decision stump algorithm on the data set to get $g$. Repeat the experiment 10000 times, each with a different data set. What is the mean of $E_{\text{out}}(g, \tau) - E_{\text{in}}(g)$ within the 10000 results? Choose the closest value.

**[a]** 0.00
**[b]** 0.02
**[c]** 0.05
**[d]** 0.30
**[e]** 0.40

**20.** (*) For $\tau = 0.1$, generate a data set of size 200 by the procedure above and run the decision stump algorithm on the data set to get $g$. Repeat the experiment 10000 times, each with a different data set. What is the mean of $E_{\text{out}}(g, \tau) - E_{\text{in}}(g)$ within the 10000 results? Choose the closest value.

**[a]** 0.00
**[b]** 0.02
**[c]** 0.05~~nbz~~
**[d]** 0.30
**[e]** 0.40