Homework #2

RELEASE DATE: 03/23/2023

DUE DATE: 04/13/2023, BEFORE 13:00 on Gradescope

QUESTIONS ARE WELCOMED ON THE NTU COOL FORUM.

You will use Gradescope to upload your choices and 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 20 problems and a total of 500 points. For each problem, there is one correct choice. If you choose the correct answer, you get 20 points; if you choose an incorrect answer, you get 0 points. For five of the secretly-selected problems, the TAs will grade your detailed solution in terms of the written explanations and/or code based on how logical/clear your solution is. Each of the five problems graded by the TAs counts as additional 20 points (in addition to the correct/incorrect choices you made). In general, each homework (except homework 0) is of a total of 500 points.

Theory of Generalization

1. A perceptron $h(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^2$ that always passes the lucky point $(11.26, 62.11)$ can be written as

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x}) \text{ such that } w_1(x_1 - 11.26) + w_2(x_2 - 62.11) = -w_0.$$

What is the growth function $m_{\mathcal{H}}(N)$ of all those lucky-point-passing perceptrons in $\mathbb{R}^2$? Choose the correct answer; prove your choice.

[a] N

[b] $2N + 4$

[c] $2N + 2$

[d] $2N$

[e] $2N - 2$

2. Consider a hypothesis set that contains 1126 perceptrons

$$h_m(\mathbf{x}) = \text{sign}(\mathbf{w}_m^T\mathbf{x}), \text{ for } m = 1, 2, \cdots, 1126$$

with $\mathbf{x} \in \mathbb{R}^{1+6210}$ (including x_0). What is the tightest upper bound on the possible VC dimension of this hypothesis set? Choose the correct answer; prove your choice.

[a] $\log_2(1126)$

[b] $\sqrt{1126}$

[c] $\log_2(1126 + 6211)$

[d] 6211

[e] $1126 + 6211$

3. Which of the following hypothesis set is of the smallest VC dimension among all choices? Choose the correct answer; explain your choice. A philosophical explanation is sufficient—there is no need to rigorously prove every case.

[a] unions of two positive intervals for $x \in \mathbb{R}$, which returns $+1$ if x is within at least one of the intervals.

[b] polynomial hypotheses of degree 3 for $x \in \mathbb{R}$, which are of the form $h(x) = \text{sign}(\sum_{i=0}^{3} w_i x^i)$

[c] the family of sine functions: $\{t \mapsto \sin(\omega t) : \omega \in \mathbb{R}\}$ for $x \in \mathbb{R}$, which return $+1$ if $x > \sin(\omega x)$.

[d] right triangles classifiers for $\mathbf{x} \in \mathbb{R}^2$, which return $+1$ is $\mathbf{x}$ is inside a right triangle whose sides adjacent to the right angle are parallel to the axes of $\mathbb{R}^2$ and with the right angle in the lower right corner

[e] axis-aligned squares classifiers for $\mathbf{x} \in \mathbb{R}^2$, which returns $+1$ if $\mathbf{x}$ is inside a square whose edges are parallel to the axes of $\mathbb{R}^2$

4. Consider a hypothesis set $\mathcal{H}$ in $\mathbb{R}^d$ containing hypothesis with $2M$ ($M > 1$) parameters. Each hypothesis $h(\mathbf{x})$ in $\mathcal{H}$ are defined by a_1,b_1,a_2,b_2,......,a_M,b_M that satisfies

- $a_1 > 0$;
- $a_m \le b_m$, for $1 \le m \le M$;
- $b_m < a_{m+1}$, for $1 \le m \le M - 1$,

with

$$h(\mathbf{x}) = \begin{cases} +1, \text{ if } a_m \le \mathbf{x}^T\mathbf{x} \le b_m \text{ for some } 1 \le m \le M \\ -1, \text{ otherwise} \end{cases}$$

What is the VC dimension of $\mathcal{H}$? Choose the correct answer; prove your choice. Note that if the TAs select this problem for human grading, a rigorous proof will get you all points, while a philosophical explanation will only get you partial points.

[a] M

[b] $2M$

[c] $2M + 1$

[d] M^2

[e] none of the other choice

5. How many of the following are **necessary** conditions for $d_{\text{vc}}(\mathcal{H}) \le d$? Choose the correct answer; state which conditions correspond to your choice 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

Linear Models

6. Consider a hypothesis set that contains hypotheses of the form $h(x) = wx$ for $x \in \mathbb{R}$. Combine the hypothesis set with the squared error function to minimize

$$E_{\text{in}}(w) = \frac{1}{N}\sum_{n=1}^{N}(h(x_n) - y_n)^2$$

on a given data set $\{(x_n, y_n)\}_{n=1}^{N}$. What is the optimal w? You can assume all denominators to be non-zero. Choose the correct answer; prove your choice.

[a] $\dfrac{\sum\limits_{n=1}^{N} y_n}{\sum\limits_{n=1}^{N} x_n}$

[b] $\dfrac{\sum\limits_{n=1}^{N} y_n x_n}{\sum\limits_{n=1}^{N} x_n^2}$

[c] $\dfrac{\sum\limits_{n=1}^{N} y_n^2}{\sum\limits_{n=1}^{N} y_n x_n}$

[d] $\sqrt{\dfrac{\sum\limits_{n=1}^{N} y_n^2}{\sum\limits_{n=1}^{N} x_n^2}}$

[e] none of the other choices

7. We use the technique of maximum likelihood to derive the error function of logistic regression. Actually, the technique is a fundamental tool in statistics for estimating the parameter from a sample. Consider a sample $\{x_1, x_2, \ldots, x_N\}$ that is independently generated from some underlying probability distribution. Furthermore, assume that all x_n are non-negative integers. Which of the following claim is **not true** about the estimate $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$? Choose the correct answer (false claim); explain your choice.

[a] Assume that the sample is generated from a Poisson distribution of parameter λ,

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Then, $\bar{x}$ is the maximum likelihood estimate of λ.

[b] Assume that the sample is generated from a unit-variance Gaussian distribution of mean parameter μ,

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}.$$

Then, $\bar{x}$ is the maximum likelihood estimate of μ.

[c] Assume that the sample is generated from a unit-scale Laplace distribution of mean parameter μ,

$$p(x) = \frac{1}{2} e^{-|x-\mu|}.$$

Then, $\bar{x}$ is the maximum likelihood estimate of μ.

[d] Assume that the sample is generated from a geometric distribution with parameter θ,

$$P(x) = (1-\theta)^{x-1}\theta$$

Then, $\frac{1}{\bar{x}}$ is the maximum likelihood estimate of θ.

[e] The claims in all other choices are correct.

8. In logistic regression, we consider the logistic hypotheses

$$h(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

to approximate the target function $f(\mathbf{x}) = P(+1 \mid \mathbf{x})$. We use the property that the hypotheses are sigmoid (s-shaped) to simplify the likelihood function and then take maximum likelihood to derive the error function E_{in}. Now, consider another family of sigmoid hypotheses, the scaled soft-sign functions

$$\tilde{h}(\mathbf{x}) = \frac{1 + \mathbf{w}^T\mathbf{x} + \mid \mathbf{w}^T\mathbf{x} \mid}{2 + 2 \mid \mathbf{w}^T\mathbf{x} \mid}.$$

Follow the same derivation steps to obtain the corresponding $\tilde{E}_{\text{in}}$ when using $\tilde{h}$ (and ignoring the case of $\mathbf{w}^T\mathbf{x} = 0$ for simplicity). What is $\nabla\tilde{E}_{\text{in}}(\mathbf{w})$? Choose the correct answer; list your derivation steps.

[a] $-\frac{1}{N}\sum_{n=1}^{N} \frac{y_n\mathbf{x}_n}{(1 + y_n\mathbf{w}^T\mathbf{x}_n + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid)(1 + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid)}$

[b] $-\frac{1}{N}\sum_{n=1}^{N} \frac{\mathbf{x}_n}{(1 + y_n\mathbf{w}^T\mathbf{x}_n + \mid \mathbf{w}^T\mathbf{x}_n \mid)(1 + \mid \mathbf{w}^T\mathbf{x}_n \mid)}$

[c] $-\frac{1}{N}\sum_{n=1}^{N} y_n\mathbf{x}_n \cdot \frac{1 - y_n\mathbf{w}^T\mathbf{x}_n + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}{2 + 2 \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}$

[d] $-\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \cdot \frac{1 - y_n\mathbf{w}^T\mathbf{x}_n + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}{2 + 2 \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}$

[e] none of the other choices

Beyond Gradient Descent

9. We discuss the gradient descent method to minimize the cross-entropy error function in class. The gradient descent method often called a first-order optimization algorithm, as we derived it using first-order Taylor's approximation for some very small $\mathbf{u} = \eta\mathbf{v}$ as introduced in class.

$$E_{\text{in}}(\mathbf{w}_t + \mathbf{u}) \approx E_{\text{in}}(\mathbf{w}_t) + \mathbf{u}^T \nabla E_{\text{in}}(\mathbf{w}_t).$$

Now, if we take the second-order approximation instead, we get

$$E_{\text{in}}(\mathbf{w}_t + \mathbf{u}) \approx E_{\text{in}}(\mathbf{w}_t) + \mathbf{u}^T \nabla E_{\text{in}}(\mathbf{w}_t) + \frac{1}{2}\mathbf{u}^T \nabla^2 E_{\text{in}}(\mathbf{w}_t)\mathbf{u},$$

where $\nabla^2 E_{\text{in}}(\mathbf{w}_t)$ is the Hessian metrix. Assume that $\nabla^2 E_{\text{in}}(\mathbf{w}_t)$ is positive definite (hence invertible), the optimal $\mathbf{u}$ is

$$\mathbf{u} = -(\nabla^2 E_{\text{in}}(\mathbf{w}_t))^{-1} \nabla E_{\text{in}}(\mathbf{w}_t).$$

Updating with $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \mathbf{u}$ is commonly called the Newton method for nonlinear optimization.

Consider linear regression, which comes with the squared error function as its E_{in}. What is the Hessian matrix of the error function? Choose the correct answer; list your derivation steps.

[a] $\frac{2}{N}\mathrm{X}\mathrm{X}^T$

[b] $\frac{2}{N}\mathrm{X}^T\mathrm{X}$

[c] $\frac{2}{N}\mathbf{y}^T\mathrm{X}\mathrm{X}^T\mathbf{y}$

[d] $\frac{2}{N}\mathrm{X}^T\mathbf{y}\mathbf{y}^T\mathrm{X}$

[e] none of the other choices

10. Continuing from the previous problem, when Newton method is used for linear regression, starting from $\mathbf{w}_0 = \mathbf{0}$, how many iterations does it take to reach the global minimum that satisfies $\nabla E_{\text{in}}(\mathbf{w}_t) = 0$? Choose the correct answer; list your derivation steps.

[a] 1

[b] d

[c] $d + 1$

[d] N

[e] none of the other choices

Decision Stumps

In class, we taught about the learning model of "positive rays." If you also include the "negative rays", that would basically make a one-dimensional perceptron model. 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. Following almost the same derivation as the model of positive rays, the growth function of the decision stump model for $x \in \mathbb{R}$ is $2N$ and the VC Dimension is 2.

11. When using the decision stump model, given $\epsilon = 0.05$ and $\delta = 0.1$, among the five choices, what is the smallest N such that the BAD probability of the VC bound is $\leq \delta$? Choose the correct answer; explain your choice.

[**a**] 100

[**b**] 1000

[**c**] 10000

[**d**] 100000

[**e**] 1000000

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 $x'_1 \leq x'_2 \leq x'_3 \leq \ldots \leq x'_N$

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

(3) return the $h_{s,\theta}$ with the minimum E_{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 dxxxxxc pxxxxxxxxxg 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 $[-0.5, +0.5]$.
- Generate y from x by $y = f(x)$ and then flip y to $-y$ with τ probability independently

12. Let $E_{\text{out}}(h, \tau)$ be the out-of-sample error of h. What is $E_{\text{out}}(h, \tau)$? Choose the correct answer; prove your choice.

[**a**] $\frac{1}{2}\min(|\theta|, 0.5)(1 - \tau) + \tau$

[**b**] $\min(|\theta|, 0.5)(1 - \tau) + \tau$

[**c**] $\frac{1}{2}\min(|\theta|, 0.5)(1 - 2\tau) + \tau$

[**d**] $\min(|\theta|, 0.5)(1 - 2\tau) + \tau$

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

13. (*) For $\tau = 0$, which means that your data is noiseless. 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? Choose the closest value; upload your source code. (*By the results in the previous problem, 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] 0.15

[b] 0.30

[c] 0.45

[d] 0.60

[e] 0.75

14. (*) For $\tau = 0$, generate a data set of size 128 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; upload your source code.

[a] 0.0020

[b] 0.0040

[c] 0.0060

[d] 0.0080

[e] 0.0100

15. (*) For $\tau = 0.20$, 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? Choose the closest value; upload your source code.

[a] 0.14

[b] 0.28

[c] 0.42

[d] 0.56

[e] 0.70

16. For $\tau = 0.20$, generate a data set of size 128 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; upload your source code.

[a] -0.0179

[b] 0.0139

[c] 0.0439

[d] 0.0739

[e] 0.1039

Decision stumps can also work for multi-dimensional data. In particular, each decision stump now deals with a specific dimension i, as shown below.

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

Implement the following decision stump algorithm for multi-dimensional data:

a) for each dimension $i = 1, 2, \cdots, d$, find the best decision stump $h_{s,i,\theta}$ using the one-dimensional decision stump algorithm that you have just implemented.

b) return the "best of best" decision stump in terms of E_{in}. If there is a tie, please choose the one with the smallest i.

The training data $\mathcal{D}$ is available at:

http://www.csie.ntu.edu.tw/~htlin/course/ml23spring/hw2/hw2_train.dat

The testing data $\mathcal{D}_{\text{test}}$ is available at:

http://www.csie.ntu.edu.tw/~htlin/course/ml23spring/hw2/hw2_test.dat

17. (*) Run the "best of best" algorithm on $\mathcal{D}$. What is E_{in} of the returned decision stump? Choose the closest value; upload your source code.

[a] 0.0065

[b] 0.0130

[c] 0.0260

[d] 0.0390

[e] 0.0780

18. (*) Use the returned decision stump to predict the label of each example within the $\mathcal{D}_{\text{test}}$ to estimate E_{out}. What is the estimated E_{out}? Choose the closest value; upload your source code.

[a] 0.0156

[b] 0.0195

[c] 0.0260

[d] 0.0391

[e] 0.0781

19. (*) Now, consider an alternative "learning" algorithm for selecting a decision stump:

b') return the "worst of best" decision stump in terms of E_{in}. That is, choose the i such that $E_{\text{in}}(h_{s,i,\theta})$ is the **largest**, where $h_{s,i,\theta}$ is the best (smallest E_{in}) decision stump on dimension i. If there is a tie, please choose the one with the smallest i.

We are curious about the difference in performance between the "best of best" $(h_{s,i^*,\theta})$ and "worst of best" $(h_{s,i^\flat,\theta})$ on $\mathcal{D}$ (for computing E_{in}) and $\mathcal{D}_{\text{test}}$ (for estimating E_{out}), i.e.,

$$\begin{aligned}\Delta E_{\text{in}} &= E_{\text{in}}(h_{s,i^\flat,\theta}) - E_{\text{in}}(h_{s,i^*,\theta})\\ \Delta E_{\text{out}} &= E_{\text{out}}(h_{s,i^\flat,\theta}) - E_{\text{out}}(h_{s,i^*,\theta}).\end{aligned}$$

What is ΔE_{in}? Choose the closest value; upload your source code.

[a] 0.00

[b] 0.10

[c] 0.20

[d] 0.30

[e] 0.40

20. (*) Continuing from the previous problem, what is the ΔE_{out}? Choose the closest value; upload your source code.

[a] 0.25

[b] 0.35

[c] 0.45

[d] 0.55

[e] 0.65