Homework #2

RELEASE DATE: 10/21/2021

DUE DATE: 11/11/2021, 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 16 problems and a total of 400 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 four 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 four 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 400 points.

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 choice.

[**a**] $\{(2,3,4),(4,3,2),(3,3,3)\}$

[**b**] $\{(1,1,1),(2,3,4),(4,3,2),(4,2,3)\}$

[**c**] $\{(1,1,1),(2,3,4),(4,3,2),(2,2,2)\}$

[**d**] $\{(1,1,1),(2,3,4),(4,3,2),(4,2,3),(3,2,4)\}$

[**e**] none of the other choices (none of them can be shattered)

2. What is the growth function of origin-passing perceptrons in 2D? Those perceptrons are all perceptrons with $w_0 = 0$. Choose the correct answer; explain your choice.

Hint: Put your input vectors on the unit circle, and perhaps think about a polar coordinate system.

[**a**] $2N + 2$

[**b**] $2N + 1$

[**c**] $2N$

[**d**] $2N - 1$

[**e**] $2N - 2$

Donut Hypothesis Set

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

$$h(\mathbf{x}) = \begin{cases} +1 & \text{if } a \le \sum_{i=1}^{d} x_i^2 \le b, \\ -1 & \text{otherwise.} \end{cases}$$

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

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

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

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

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

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

4. Following the previous problem, what is the VC dimension of the donut hypothesis set? Choose the correct answer; explain your choice.

[**a**] d

[**b**] 6

[**c**] 3

[**d**] 2

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

More on VC Dimension

5. Which of the following hypothesis set is of a different VC dimension, compared to others? Choose the correct answer; explain your choice.

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

[**b**] Axis-aligned rectangle classifiers for $\mathbf{x} \in \mathbb{R}^2$, which returns $+1$ if $\mathbf{x}$ is inside a rectangle whose edges are parallel to the axes of $\mathbb{R}^2$.

[**c**] Positively-biased perceptrons over $\mathbf{x} \in \mathbb{R}^4$, which contains perceptrons with $w_0 > 0$.

[**d**] Polynomial hypotheses of degree 3 for $x \in \mathbb{R}$, which are of the form

$$h(x) = \text{sign}(\sum_{i=0}^{3} w_i x^i).$$

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

6. For a finite hypothesis set $\mathcal{H}$ with 1126 binary classifiers, what is the largest possible value of $d_{\text{vc}}(\mathcal{H})$? Choose the correct answer; explain your choice.

[a] 1126

[b] 112

[c] 11

[d] 10

[e] 1

Deviation from Optimal Hypothesis

7. Recall that the multiple-bin Hoeffding bound quantifies the BAD probability from *any* hypothesis h in the hypothesis set. That is,

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \le 2M \exp\left(-2\epsilon^2 N\right).$$

Define the best-E_{in} hypothesis

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

and the best-E_{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 multiple-bin Hoeffding 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 choice.

[a] $\sqrt{\frac{1}{2N}\ln\left(\frac{M}{\delta}\right)}$

[b] $\sqrt{\frac{1}{N}\ln\left(\frac{2M}{\delta}\right)}$

[c] $\sqrt{\frac{1}{2N}\ln\left(\frac{2M}{\delta}\right)}$

[d] $2\sqrt{\frac{1}{2N}\ln\left(\frac{2M}{\delta}\right)}$

[e] $\sqrt{\frac{1}{N}\ln\left(\frac{M}{\delta}\right)}$

VC Bound

8. When using the positive ray model taught in class, 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

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

is $\le \delta$? Choose the correct answer; explain your choice.

[a] 10000

[b] 11000

[c] 12000

[d] 13000

[e] 14000

Hessian and Newton Method

9. Let $E(\mathbf{w})\colon \mathbb{R}^d \to \mathbb{R}$ be a function. Denote the gradient $\mathbf{b}_E(\mathbf{w})$ and the Hessian $\mathrm{A}_E(\mathbf{w})$ by

$$\mathbf{b}_E(\mathbf{w}) = \nabla E(\mathbf{w}) = \begin{bmatrix} \frac{\partial E}{\partial w_1}(\mathbf{w}) \\ \frac{\partial E}{\partial w_2}(\mathbf{w}) \\ \vdots \\ \frac{\partial E}{\partial w_d}(\mathbf{w}) \end{bmatrix}_{d\times 1} \quad \text{and } \mathrm{A}_E(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_1 \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_d}(\mathbf{w}) \\ \frac{\partial^2 E}{\partial w_2 \partial w_1}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_2^2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_2 \partial w_d}(\mathbf{w}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_d \partial w_1}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_d \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_d^2}(\mathbf{w}) \end{bmatrix}_{d\times d}.$$

Then, the second-order Taylor's expansion of $E(\mathbf{w})$ around $\mathbf{u}$ is:

$$E(\mathbf{w}) \approx E(\mathbf{u}) + \mathbf{b}_E(\mathbf{u})^T(\mathbf{w}-\mathbf{u}) + \frac{1}{2}(\mathbf{w}-\mathbf{u})^T \mathrm{A}_E(\mathbf{u})(\mathbf{w}-\mathbf{u}).$$

Suppose $\mathrm{A}_E(\mathbf{u})$ is positive definite. What is the optimal direction $\mathbf{v}$ such that $\mathbf{w} \leftarrow \mathbf{u}+\mathbf{v}$ minimizes the right-hand-side of the Taylor's expansion above? Choose the correct answer; explain your choice.

Note that iterative optimization with $\mathbf{v}$ *is generally called Newton's method.*

[a] $+(\mathrm{A}_E(\mathbf{u}))^{-1}\mathbf{b}_E(\mathbf{u})$

[b] $-(\mathrm{A}_E(\mathbf{u}))^{-1}\mathbf{b}_E(\mathbf{u})$

[c] $+(\mathrm{A}_E(\mathbf{u}))^{+1}\mathbf{b}_E(\mathbf{u})$

[d] $-(\mathrm{A}_E(\mathbf{u}))^{+1}\mathbf{b}_E(\mathbf{u})$

[e] none of the other choices

10. Following the previous problem, considering minimizing $E_{\text{in}}(\mathbf{w})$ in logistic regression problem with Newton's method. Consider a data set $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ with the cross-entropy error function for E_{in}:

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n \mathbf{w}^T \mathbf{x}_n))$$

For any given $\mathbf{w}_t$, let

$$h_t(\mathbf{x}) = \frac{1}{1+\exp(\mathbf{w}_t^T \mathbf{x})}.$$

What is the Hessian $\mathrm{A}_E(\mathbf{w}_t)$ with $E = E_{\text{in}}$? Choose the correct answer; explain your choice.

[a] $\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$

[b] $\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{w}_t\|^2 \mathbf{x}_n \mathbf{x}_n^T$

[c] $\frac{1}{N}\sum_{n=1}^{N} h_t^2(y_n \mathbf{x}_n) \mathbf{x}_n \mathbf{x}_n^T$

[d] $\frac{1}{N}\sum_{n=1}^{N} h_t(y_n \mathbf{x}_n) h_t(-y_n \mathbf{x}_n) \mathbf{x}_n \mathbf{x}_n^T$

[e] none of the other choices

Linear Regression

11. In the lecture, we learned that the linear regression weights $\mathbf{w}$ must satisfy the normal equation $\mathrm{X}^T\mathrm{X}\mathbf{w} = \mathrm{X}^T\mathbf{y}$. If $\mathrm{X}^T\mathrm{X}$ is not invertible, we cannot compute $\mathbf{w}_{\text{LIN}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathbf{y}$. Then, one possible solution is

$$\mathbf{w}_{\text{LIN}} = \mathrm{X}^\dagger\mathbf{y},$$

where $\mathrm{X}^\dagger$ is called the Moore-Penrose pseudo-inverse. Let $\mathrm{X} = \mathrm{U}\Sigma\mathrm{V}^T$ be the singular value decomposition of the N by $d+1$ matrix X, with U and V being unitary matrices. The Moore-Penrose pseudo-inverse $\mathrm{X}^\dagger = \mathrm{V}\Sigma^\dagger\mathrm{U}^T$ is a $d+1$ by N matrix, where $\Sigma^\dagger[i][n] = \frac{1}{\Sigma[n][i]}$ when $\Sigma[n][i]$ is nonzero, and 0 otherwise. Which of the following statements related to $\mathrm{X}^\dagger$ is incorrect? Choose the **incorrect** statement; explain your choice.

[**a**] $(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T = \mathrm{X}^\dagger$ when $\mathrm{X}^T\mathrm{X}$ is invertible.

[**b**] For any $k \in \mathbb{Z}^+$, $(\mathrm{X}\mathrm{X}^\dagger)^k = \mathrm{X}\mathrm{X}^\dagger$.

[**c**] $\mathrm{X}\mathrm{X}^\dagger = \mathrm{I}_N$, the N by N identity matrix.

[**d**] $\mathrm{trace}(\mathrm{X}\mathrm{X}^\dagger) = \mathrm{rank}(\mathrm{X})$.

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

12. In the lecture, we learned that the logistic regression can be derived by maximizing the likelihood function where each label y_n is generated from $P(+1|\mathbf{x}_n)$ "pretended by" $1/(1 + \exp(-\mathbf{w}^T\mathbf{x}_n))$. Now, consider a case where each real-valued label y_n is generated from $p(y|\mathbf{x}_n)$, where p is used instead of P to denote a probability density function (instead of a probability mass function—you can ignore the subtle mathematical difference for now). We will consider $p(y|\mathbf{x}_n)$ to be pretended by a normal distribution with mean $\mathbf{w}^T\mathbf{x}_n$ and variance a^2. Assume that all inversions in the equations below exist, what is the optimal $\mathbf{w}^*$ that maximizes the likelihood function for this case? Choose the correct answer; explain your choice.

[**a**] $\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathbf{y}$

[**b**] $a\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathbf{y}$

[**c**] $a^2\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathbf{y}$

[**d**] $\left(\mathrm{X}^T\mathrm{X} + a^2\mathrm{I}\right)^{\dagger}\mathrm{X}^T\mathbf{y}$

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

Experiments with Linear Models

Next, we will play with linear regression, logistic regression, and their use for binary classification. We will also learn how the their different robustness to outliers. We will generate artificial $2D$ data for training and testing in the next problems. Each example $(\mathbf{x}, y)$ is assumed to be generated from the following process:

- Flip a fair coin to get either $y = +1$ or $y = -1$.
- If $y = +1$, generate $\mathbf{x} = (1, x_1, x_2)$ where (x_1, x_2) comes from a normal distribution of mean $[2, 3]$ and covariance $\begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$.
- If $y = -1$, generate $\mathbf{x} = (1, x_1, x_2)$ where (x_1, x_2) comes from a normal distribution of mean $[0, 4]$ and covariance $\begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}$.

Please generate $N = 200$ examples from the process as your training data set $\mathcal{D}$. Then, generate 5000 more examples from the process as your test data set (for evaluating E_{out}).

Hint: Be sure to check whether your normal distribution function needs you to provide the variance, which would be like 0.6 *for the* $y_n = +1$ *cases, or the standard deviation, which would be like* $\sqrt{0.6}$.

13. (*) Implement the linear regression algorithm taught in the lecture. Run the algorithm for 100 times, each with a different random seed for generating the two data sets above. What is the average $E_{\text{in}}^{\text{sqr}}(\mathbf{w}_{\text{lin}})$, where $E_{\text{in}}^{\text{sqr}}$ denotes the *averaged* squared error over N examples? Choose the closest answer; provide your code.

[**a**] 0.04

[**b**] 0.16

[**c**] 0.24

[**d**] 0.28

[**e**] 0.40

14. (*) Following the previous problem, what is the average $\left|E_{\text{in}}^{0/1}(\mathbf{w}_{\text{lin}}) - E_{\text{out}}^{0/1}(\mathbf{w}_{\text{lin}})\right|$, where 0/1 denotes the 0/1 error (i.e. using $\mathbf{w}_{\text{lin}}$ for binary classification), $E_{\text{in}}^{0/1}$ denotes the *averaged* 0/1 error over N examples, $E_{\text{out}}^{(0/1)}$ is estimated using the averaged 0/1 error on the test data set above? Choose the closest answer; provide your code.

[**a**] 0.091

[**b**] 0.065

[**c**] 0.039

[**d**] 0.013

[**e**] 0.001

15. (*) Consider two algorithms. The first one, $\mathcal{A}$, is the linear regression algorithm above. The second one $\mathcal{B}$ is logistic regression, trained with gradient descent with $\eta = 0.1$ for $T = 500$ iterations, starting from $\mathbf{w}_0 = \mathbf{0}$. Run the algorithms on the same $\mathcal{D}$, and record $[E_{\text{out}}^{0/1}(\mathcal{A}(\mathcal{D})), E_{\text{out}}^{0/1}(\mathcal{B}(\mathcal{D}))]$. Repeat the process for 100 times, each with a different random seed for generating the two data sets above. What is the average $[E_{\text{out}}^{0/1}(\mathcal{A}(\mathcal{D})), E_{\text{out}}^{0/1}(\mathcal{B}(\mathcal{D}))]$? Choose the closest answer; provide your code.

[**a**] (0.018, 0.018)

[**b**] (0.058, 0.058)

[**c**] (0.058, 0.093)

[**d**] (0.138, 0.108)

[**e**] (0.268, 0.208)

16. (*) Following the previous problem, in addition to the 200 examples in $\mathcal{D}$, add 20 outlier examples generated from the following process to your training data (but not to your test data). All outlier examples will be labeled $y = +1$ and $\mathbf{x} = [1, x_1, x_2]$ where (x_1, x_2) comes from a normal distribution of mean $[6, 0]$ and covariance $\begin{bmatrix} 0.3 & 0 \\ 0 & 0.1 \end{bmatrix}$. Name the new training data set $\mathcal{D}'$. Run the algorithms on the same $\mathcal{D}'$, and record $[E_{\text{out}}^{0/1}(\mathcal{A}(\mathcal{D}')), E_{\text{out}}^{0/1}(\mathcal{B}(\mathcal{D}'))]$. Repeat the process for 100 times, each with a different random seed for generating the two data sets above. What is the average $[E_{\text{out}}^{0/1}(\mathcal{A}(\mathcal{D}')), E_{\text{out}}^{0/1}(\mathcal{B}(\mathcal{D}'))]$? Choose the closest answer; provide your code.

[**a**] (0.070, 0.018)

[**b**] (0.240, 0.048)

[**c**] (0.090, 0.058)

[**d**] (0.090, 0.078)

[**e**] (0.270, 0.108)