Homework #3

RELEASE DATE: 10/07/2024

RED CORRECTION: 10/12/2024 16:30

DUE DATE: 10/21/2024, BEFORE 13:00 on GRADESCOPE

QUESTIONS ARE WELCOMED ON DISCORD (INFORMALLY) OR VIA EMAILS (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 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 200 points and 20 bonus points. In general, every homework set would come with a full credit of 200 points, with some possible bonus points.

1. (10 points, auto-graded) Which of the following hypothesis set, each parameterized by one parameter only, is of the largest d_{vc}?

[a] $\{c_s(\mathbf{x})\colon s \in \{-1, +1\}\}$ where $c_s(\mathbf{x}) = s$

[b] $\{r_\theta(\mathbf{x})\colon \theta \in \mathbb{R}\}$ where $r_\theta(\mathbf{x}) = \text{sign}(x_1 - \theta)$

[c] $\{q_i(\mathbf{x})\colon i \in \{1, 2, \dots, d\}\}$ where $q_i(\mathbf{x}) = \text{sign}(x_i)$

[d] $\{u_\alpha(\mathbf{x})\colon \alpha \in \mathbb{R}\}$ where $u_\alpha(\mathbf{x}) = \text{sign}(\sin(\alpha x_1))$

[e] $\{v_\beta(\mathbf{x})\colon \beta \in \mathbb{R}\}$ where $v_\beta(\mathbf{x}) = \text{sign}(\beta x_1)$

2. (10 points, auto-graded) 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_{LIN}? You can assume all denominators to be non-zero. *(Hint: This is linear regression in $\mathbb{R}$ without the added x_0.)*

[a] $\frac{\sum_{n=1}^N x_n x_n}{\sum_{n=1}^N x_n y_n}$

[b] $\frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n x_n}$

[c] $\frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N y_n y_n}$

[d] $\frac{\sum_{n=1}^N y_n y_n}{\sum_{n=1}^N x_n y_n}$

[e] none of the other choices

3. (10 points, auto-graded) In Lecture 9, we introduced the hat matrix $\mathrm{H} = \mathrm{X}\mathrm{X}^\dagger$ for linear regression. The matrix projects the label vector $\mathbf{y}$ to the "predicted" vector $\hat{\mathbf{y}} = \mathrm{H}\mathbf{y}$ and helps us analyze the error of linear regression. Assume that $\mathrm{X}^T\mathrm{X}$ is invertible, which makes $\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$. Now, consider the following operations on X. Which operation can possibly change H?

[a] multiplying each of the n-th row of X by $\frac{1}{n}$ (which is equivalent to scaling the n-th example by $\frac{1}{n}$)

[b] multiplying each of the i-th column of X by i^2 (which is equivalent to scaling the i-th feature by i^2)

[c] multiplying the whole matrix X by 2 (which is equivalent to scaling all input vectors by 2)

[d] adding three randomly-chosen columns i, j, k to column 1 of X
(i.e., $x_{n,1} \leftarrow x_{n,1} + x_{n,i} + x_{n,j} + x_{n,k}$)

[e] none of the other choices (i.e. all other choices are guaranteed to keep H unchanged.)

4. (10 points, auto-graded) Let $y_1, y_2, \ldots, y_N$ be N values generated i.i.d. from a uniform distribution $[\theta, 1]$ with some unknown θ. For any $\hat{\theta} \leq \min(y_1, y_2, \ldots, y_N)$, what is its likelihood?

[a] $\left(\frac{1}{\hat{\theta}}\right)^N$

[b] $\prod_{n=1}^{N} \frac{y_n}{1-\hat{\theta}}$

[c] $\left(\frac{1}{1-\hat{\theta}}\right)^N$

[d] $\frac{\max(y_1,\ldots,y_N)}{\hat{\theta}}$

[e] $\frac{\min(y_1,\ldots,y_N)}{1-\hat{\theta}}$

(Hint: Those who are interested in more math [who isn't? :-)] are encouraged to try to derive the maximum-likelihood estimator.)

5. (20 points, human-graded) Prove or disprove that for any two non-empty hypothesis sets $\mathcal{H}_1$ and $\mathcal{H}_2$ for binary classification that operate on the same input space, $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_{vc}(\mathcal{H}_1) + d_{vc}(\mathcal{H}_2)$. Note that the $\cup$ operation represents set-union. That is, $\{h_1, h_2, h_3\} \cup \{h_2, h_4\} = \{h_1, h_2, h_3, h_4\}$.

6. (20 points, human-graded) Consider a binary classification problem, where $\mathcal{Y} = \{-1, +1\}$. Assume a noisy scenario where the data is generated i.i.d. from some $P(\mathbf{x}, y)$. In class, we discussed that when the 0/1 error function (i.e. classification error) is considered, calculating the "ideal mini-target" on each $\mathbf{x}$ reveals the hidden target function of

$$f_{0/1}(\mathbf{x}) = \mathrm{argmax}_{y \in \{-1,+1\}} P(y|\mathbf{x}) = \mathrm{sign}\left(P(y = +1|\mathbf{x}) - \frac{1}{2}\right).$$

Instead of the 0/1 error, if we consider the super-market error function, where a false negative (classifying a positive example as a negative one) is 10 times more important than a false positive, the hidden target should be changed to

$$f_{\mathrm{MKT}}(\mathbf{x}) = \mathrm{sign}\big(P(y = +1|\mathbf{x}) - \alpha\big).$$

Prove what the value of α should be.

7. (20 points, human-graded) In class, we had two definitions of $E_{\text{out}}(h)$ for binary classification. The first definition compares the hypothesis h against the target function f.

$$E_{\text{out}}^{(1)}(h) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})} [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!] .$$

The second definition extends from the first definition, and compares the hypothesis h against the noisy distribution $P(y \mid \mathbf{x})$.

$$E_{\text{out}}^{(2)}(h) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}),y\sim P(y|\mathbf{x})} [\![h(\mathbf{x}) \neq y]\!] .$$

Note that when considering the 0/1 error, we know that the target function $f(\mathbf{x})$ hides itself within $P(y \mid \mathbf{x})$ by

$$f(\mathbf{x}) = \text{argmax}_{y\in\{-1,+1\}} P(y|\mathbf{x}) = \text{sign}\left(P(y=+1|\mathbf{x}) - \frac{1}{2}\right).$$

With all the definitions above, prove that for any hypothesis h,

$$E_{\text{out}}^{(2)}(h) \leq E_{\text{out}}^{(1)}(h) + E_{\text{out}}^{(2)}(f).$$

(Hint: Technically, $E_{out}^{(2)}(f)$ is a constant that represents the irreducible error (i.e. noise) of the learning problem.)

8. (20 points, human-graded) Consider running linear regression on $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where $\mathbf{x}_n$ includes the constant dimension $x_0 = 1$ as usual. For simplicity, you can assume that $\mathrm{X}^T\mathrm{X}$ is invertible. Assume that the unique (why :-)) solution $\mathbf{w}_{\text{LIN}}$ is obtained after running linear regression on the data above. Then, if every x_0 is changed to 1126 instead of 1, run linear regression again to get the unique solution $\mathbf{w}_{\text{LUCKY}}$. Prove that $\mathbf{w}_{\text{LIN}} = \mathrm{D}\mathbf{w}_{\text{LUCKY}}$, where D is some diagonal matrix, by deriving the correct D.

9. (20 points, human-graded) 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,

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

Follow the same derivation steps to obtain the corresponding $\tilde{E}_{\text{in}}$ when using $\tilde{h}$. What is $\nabla\tilde{E}_{\text{in}}(\mathbf{w})$?

10. (20 points, code needed, human-graded) Next, we use a real-world data set to study linear regression. Please download the `cpusmall_scale` data set at

`https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression/cpusmall_scale`

We use the column-scaled version instead of the original version so you'd likely encounter fewer numerical issues.

The data set contains 8192 examples. In each experiment, you are asked to

(1) randomly sample N out of 8192 examples as your training data.

(2) add $x_0 = 1$ to each example, as always.

(3) run linear regression on the N examples, using any reasonable implementations of $\mathrm{X}^\dagger$ on the input matrix X, to get $\mathbf{w}_{lin}$

(4) evaluate $E_{\text{in}}(\mathbf{w}_{lin})$ by averaging the squared error over the N examples; estimate $E_{\text{out}}(\mathbf{w}_{lin})$ by averaging the squared error over the rest $(8192 - N)$ examples

For $N = 32$, run the experiment above for 1126 times, and plot a scatter plot of $(E_{\text{in}}(\mathbf{w}_{lin}), E_{\text{out}}(\mathbf{w}_{lin}))$ in each experiment. Describe your findings.

Then, provide the first page of the snapshot of your code as a proof that you have written the code.

11. (20 points, code needed, human-graded) For each of $N = 25, 50, 75, 100, \ldots, 2000$, calculate $\bar{E}_{in}(N)$ and $\bar{E}_{out}(N)$ by averaging E_{in} and E_{out} over 16 experiments. Then, plot the learning curves that show $\bar{E}_{in}(N)$ and $\bar{E}_{out}(N)$ as functions of N on the same figure. Describe your findings.

Then, provide the first page of the snapshot of your code as a proof that you have written the code.

12. (20 points, code needed, human-graded) Repeat Problem 11, but using the first 2 features for each example instead of all 12 features. That is, run linear regression with $\mathbf{x} = [x_0, x_1, x_2]$ instead. Describe your findings. In particular, compare your results here to those of Problem 11.

Then, provide the first page of the snapshot of your code as a proof that you have written the code.

13. (Bonus 20 points, human graded) Please note that this part is related to the "optional" lecture 6 of the course. If you want to get the bonus, you need to do something "extra" to at least understand the definitions below. We hope that this reminds everyone that you do not always need to solve the bonus problem! In Lecture 6, we proved $B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$. Now, prove that $B(N,k) \geq \sum_{i=0}^{k-1} \binom{N}{i}$. Thus, $B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$.