HW2 Choices

Graded

Student

Chih-Hao Liao

Total Points

380 / 400 pts

Question 1

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 2

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 3

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 4

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 5

(no title) **0** / 20 pts

+ 20 pts Correct

✔ **+ 0 pts** Incorrect

Question 6

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 7

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 8

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 9

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 10

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 11

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 12

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 13

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 14

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 15

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 16

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 17

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 18

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 19

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 20

(no title) **20** / 20 pts

✔ **+ 20 pts** Correct

+ 0 pts Incorrect

Question 21

(no title) **0** / 0 pts

+ 0 pts Incorrect

✔ **+ 0 pts** Correct

Q1
20 Points

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.

- ○ N
- ○ $2N + 4$
- ○ $2N + 2$
- ◉ $2N$
- ○ $2N - 2$

Q2
20 Points

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.

- ◉ $\log_2(1126)$
- ○ $\sqrt{1126}$
- ○ $\log_2(1126 + 6211)$
- ○ 6211
- ○ $1126 + 6211$

Q3

20 Points

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.

- ○ unions of two positive intervals for $x \in \mathbb{R}$, which returns $+1$ if x is within at least one of the intervals.
- ○ polynomial hypotheses of degree 3 for $x \in \mathbb{R}$, which are of the form $h(x) = \text{sign}(\sum_{i=0}^{3} w_i x^i)$
- ○ the family of sine functions: $\{t \mapsto \sin(\omega t) : \omega \in \mathbb{R}\}$ for $x \in \mathbb{R}$, which return $+1$ if $x > \sin(\omega x)$.
- ○ right triangles classifiers for $\mathbf{x} \in \mathbb{R}^2$, which return $+1$ if $\mathbf{x}$ 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
- ◉ 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 $\mathbb{R}^2$

Q4
20 Points

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.

- ○ M
- ◉ $2M$
- ○ $2M+1$
- ○ M^2
- ○ none of the other choice

Q5

20 Points

How many of the following are **necessary** conditions for $d_{\mathrm{vc}}(\mathcal{H}) \le d$? 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 $\mathcal{H}$

- ○ 1
- ○ 2
- ◉ 3
- ○ 4
- ○ 5

Q6
20 Points

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.

- ○ $\dfrac{\sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n}$
- ● $\dfrac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}$
- ○ $\dfrac{\sum_{n=1}^{N} y_n^2}{\sum_{n=1}^{N} y_n x_n}$
- ○ $\sqrt{\dfrac{\sum_{n=1}^{N} y_n^2}{\sum_{n=1}^{N} x_n^2}}$
- ○ none of the other choices

Q7

20 Points

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 \textbf{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.

- ○ 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 .
- ○ 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.
- ◉ 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.
- ○ 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.
- ○ The claims in all other choices are correct.

Q8

20 Points

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.

- (selected) $-\frac{1}{N}\sum_{n=1}^{N} \frac{y_n\mathbf{x}_n}{(1 + y_n\mathbf{w}^T\mathbf{x}_n + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid)(1 + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid)}$
- $-\frac{1}{N}\sum_{n=1}^{N} \frac{\mathbf{x}_n}{(1 + y_n\mathbf{w}^T\mathbf{x}_n + \mid \mathbf{w}^T\mathbf{x}_n \mid)(1 + \mid \mathbf{w}^T\mathbf{x}_n \mid)}$
- $-\frac{1}{N}\sum_{n=1}^{N} y_n\mathbf{x}_n \cdot \frac{1 - y_n\mathbf{w}^T\mathbf{x}_n + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}{2 + 2 \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}$
- $-\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \cdot \frac{1 - y_n\mathbf{w}^T\mathbf{x}_n + \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}{2 + 2 \mid y_n\mathbf{w}^T\mathbf{x}_n \mid}$
- none of the other choices

Q9

20 Points

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.

- ○ $\frac{2}{N}\mathrm{X}\mathrm{X}^T$
- ◉ $\frac{2}{N}\mathrm{X}^T\mathrm{X}$
- ○ $\frac{2}{N}\mathbf{y}^T\mathrm{X}\mathrm{X}^T\mathbf{y}$
- ○ $\frac{2}{N}\mathrm{X}^T\mathbf{y}\mathbf{y}^T\mathrm{X}$
- ○ none of the other choices

Q10

20 Points

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.

- (●) 1
- ( ) d
- ( ) $d+1$
- ( ) N
- ( ) none of the other choices

Q11

20 Points

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.

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.

- ( ) 100
- ( ) 1000
- ( ) 10000
- (●) 100000
- ( ) 1000000

Q12

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 $\theta \in \{-\infty\} \cup \{\frac{x'_i + x'_{i+1}}{2} : 1 \le i \le N-1 \text{ and } x'_i \ne x'_{i+1}\}$ 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 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

Let $E_{\text{out}}(h, \tau)$ be the out-of-sample error of $h(+1, \theta)$. What is $E_{\text{out}}(h, \tau)$? Choose the correct answer; prove your choice.

- ○ $\frac{1}{2}\min(|\theta|, 0.5)(1-\tau) + \tau$
- ○ $\min(|\theta|, 0.5)(1-\tau) + \tau$
- ○ $\frac{1}{2}\min(|\theta|, 0.5)(1-2\tau) + \tau$
- ◉ $\min(|\theta|, 0.5)(1-2\tau) + \tau$
- ○ none of the other choices

Q13
20 Points

(*) 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.
({\it 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}).

- ○ 0.15
- ◉ 0.30
- ○ 0.45
- ○ 0.60
- ○ 0.75

Q14
20 Points

(*) 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.

- ○ 0.0020
- ◉ 0.0040
- ○ 0.0060
- ○ 0.0080
- ○ 0.0100

Q15

20 Points

(*) 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.

- ○ 0.14
- ○ 0.28
- ◉ 0.42
- ○ 0.56
- ○ 0.70

Q16

20 Points

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.

- ○ -0.0179
- ◉ 0.0139
- ○ 0.0439
- ○ 0.0739
- ○ 0.1039

Q17

20 Points

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:

1. 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.
2. 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

The files are of the same format as the one you received in Homework 1.

(*) 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.

- ○ 0.0065
- ○ 0.0130
- ◉ 0.0260
- ○ 0.0390
- ○ 0.0780

Q18

20 Points

(*) 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.

- ○ 0.0156
- ○ 0.0195
- ○ 0.0260
- ○ 0.0391
- ◉ 0.0781

Q19

20 Points

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

2. 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 \textbf{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.

- ○ 0.00
- ○ 0.10
- ○ 0.20
- ◉ 0.30
- ○ 0.40

Q20
20 Points

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

- ○ 0.25
- ◉ 0.35
- ○ 0.45
- ○ 0.55
- ○ 0.65

Q21
0 Points

How many gold medals do you want to use for this homework (every gold medal extends the deadline of this homework by 12 hours, and you have four gold medals in total this semester)

- ◉ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6