HW4 Choices

Graded

Student

Chih-Hao Liao

Total Points

180 / 180 pts

Question 1

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

+ **0 pts** Incorrect

✔ + **0 pts** Correct

Question 2

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

✔ + **0 pts** Correct

+ **0 pts** Incorrect

Question 3

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

✔ + **0 pts** Correct

+ **0 pts** Incorrect

Question 4

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

✔ + **0 pts** Correct

+ **0 pts** Incorrect

Question 5

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

✔ + **0 pts** Correct

+ **0 pts** Incorrect

Question 6

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

✔ + **0 pts** Correct

+ **0 pts** Incorrect

Question 7

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

✔ **+ 0 pts** Correct

+ 0 pts Incorrect

Question 8

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

✔ **+ 0 pts** Correct

+ 0 pts Incorrect

Question 9

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

✔ **+ 0 pts** Correct

+ 0 pts Incorrect

Question 10

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

✔ **+ 0 pts** Correct

+ 0 pts Incorrect

Question 11

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

✔ **+ 0 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** Correct

+ 0 pts Incorrect

Q1

0 Points

Consider a one-dimensional data set $\{(x_n, y_n)\}_{n=1}^{N}$ where each $x_n \in \mathbb{R}$ and $y_n \in \mathbb{R}$. Then, solve the following one-variable regularized linear regression problem:

$$\min_{w \in \mathbb{R}} \frac{1}{N} \sum_{n=1}^{N} (w \cdot x_n - y_n)^2 + \frac{\lambda}{N} w^2.$$

If the optimal solution to the problem above is w^*, it can be shown that w^* is also the optimal solution of

$$\min_{w \in \mathbb{R}} \frac{1}{N} \sum_{n=1}^{N} (w \cdot x_n - y_n)^2 \text{ subject to } w^2 \leq C$$

with $C = (w^*)^2$. This allows us to express the relationship between C in the constrained optimization problem and λ in the augmented optimization problem for any $\lambda > 0$. What is the relationship?
Choose the correct answer; explain your answer.

note: All the choices hint you that a smaller λ corresponds to a bigger C.

- [x] $$C = \left(\frac{\sum_{n=1}^{N} x_n y_n}{\sum_{n=1}^{N} x_n^2 + \lambda} \right)^2$$
- [ ] $$C = \left(\frac{\sum_{n=1}^{N} y_n^2}{\sum_{n=1}^{N} x_n^2 + \lambda} \right)^2$$
- [ ] $$C = \left(\frac{\sum_{n=1}^{N} x_n^2 y_n^2}{\sum_{n=1}^{N} x_n^2 + \lambda} \right)^2$$
- [ ] $$C = \left(\frac{\sum_{n=1}^{N} x_n y_n}{\sum_{n=1}^{N} y_n^2 + \lambda} \right)^2$$
- [ ] $$C = \left(\frac{\sum_{n=1}^{N} x_n^2}{\sum_{n=1}^{N} y_n^2 + \lambda} \right)^2$$

Q2

0 Points

The ranges of features may affect regularization. One common technique to align the ranges of features is to consider a ''normalization'' transformation. Define

$\Phi(\mathbf{x}) = \Gamma^{-1}(\mathbf{x} - \mathbf{u})$, where $\mathbf{u}$ is an estimated mean of the examples, Γ is a diagonal matrix with positive diagonal values $\gamma_0, \gamma_1, \ldots, \gamma_d$ that indicate the estimated standard deviation. For simplicity, consider $\mathbf{u} = \mathbf{0}$. Then, conducting L2-regularized linear regression in the $\mathcal{Z}$-space

$$\min_{\tilde{\mathbf{w}} \in \mathbb{R}^{d+1}} \frac{1}{N} \sum_{n=1}^{N} (\tilde{\mathbf{w}}^T \Phi(\mathbf{x_n}) - y_n)^2 + \frac{\lambda}{N} (\tilde{\mathbf{w}}^T \tilde{\mathbf{w}})$$

is equivalent to regularized linear regression in the $\mathcal{X}$-space

$$\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x_n} - y_n)^2 + \frac{\lambda}{N} \Omega(\mathbf{w})$$

with a different regularizer $\Omega(\mathbf{w})$. What is $\Omega(\mathbf{w})$? Choose the correct answer; explain your answer.

- [ ] $\mathbf{w}^T \Gamma \mathbf{w}$
- [x] $\mathbf{w}^T \Gamma^2 \mathbf{w}$
- [ ] $\mathbf{w}^T \mathbf{w}$
- [ ] $\mathbf{w}^T \Gamma^{-2} \mathbf{w}$
- [ ] $\mathbf{w}^T \Gamma^{-1} \mathbf{w}$

Q3

0 Points

The error function of logistic regression

$$\mathrm{err}(\mathbf{w}, \mathbf{x}, y) = \ln(1 + \exp(-y\mathbf{w}^T\mathbf{x}))$$

can be re-written as

$$\mathrm{err}(\mathbf{w}, \mathbf{x}, y) = [\![y = +1]\!] \ln(1 + \exp(-\mathbf{w}^T\mathbf{x})) + [\![y = -1]\!] \ln(1 + \exp(\mathbf{w}^T\mathbf{x})).$$

Label smoothing is a popular way of combatting overfitting by replacing the error function with a smoothed one

$$\mathrm{err}_{smooth}(\mathbf{w}, \mathbf{x}, +1) = (1 - \frac{\alpha}{2}) \ln(1 + \exp(-\mathbf{w}^T\mathbf{x})) + \frac{\alpha}{2} \ln(1 + \exp(\mathbf{w}^T\mathbf{x})).$$

and

$$\mathrm{err}_{smooth}(\mathbf{w}, \mathbf{x}, -1) = \frac{\alpha}{2} \ln(1 + \exp(-\mathbf{w}^T\mathbf{x})) + (1 - \frac{\alpha}{2}) \ln(1 + \exp(\mathbf{w}^T\mathbf{x})).$$

Solving the in-sample error using the smoothed error function

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \mathrm{err}_{smooth}(\mathbf{w}, \mathbf{x_n}, y_n)$$

is equivalent to solving a regularized logistic regression problem.

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \mathrm{err}(\mathbf{w}, \mathbf{x_n}, y_n) + \frac{\lambda}{N} \sum_{n=1}^{N} \Omega(\mathbf{x}, \mathbf{x_n}).$$

Let $D_{KL}(P||Q)$ denote the KL-divergence between two probability distributions P and Q and let $P_u(+1) = P_u(-1) = \frac{1}{2}$ denote a uniform probability distribution on binary outcomes. Note that every logistic hypothesis $h(\mathbf{x})$ defines a probability distribution $P_h(+1|\mathbf{x}) = h(\mathbf{x})$ and $P_h(-1|\mathbf{x}) = (1 - h(\mathbf{x}))$. Let $\lambda = \frac{\alpha}{1-\alpha}$. What is $\Omega(\mathbf{w}, \mathbf{x})$? Choose the correct answer; explain your answer.

- [x] $D_{KL}(P_u||P_h)$
- [ ] $D_{KL}(P_h||P_u)$
- [ ] $\frac{1}{2}(D_{KL}(P_u||P_h) + D_{KL}(P_h||P_u))$
- [ ] $D_{KL}(P_u||P_h) + D_{KL}(P_h||P_u)$
- [ ] none of the other choices

Q4

0 Points

Consider three examples (x_1, y_1), (x_2, y_2), $(x_3, y_3 = 1)$. Assume that x_1, x_2, x_3 are independent random variables that are uniformly generated between $[-1, 1]$, and y_1, y_2 are independent random variables that are uniformly generated between $[0, 2]$. Use leave-one-out cross-validation with the squared error to estimate the performance of the constant model, which returns the best constant hypothesis $h(x) = w_0$ in terms of the squared error. What is the probability that $E_{loocv} \leq \frac{1}{3}$? Choose the correct answer; explain your choice.

- [ ] $\frac{\pi}{12}$
- [x] $\frac{\pi}{3\sqrt{3}}$
- [ ] $\frac{\pi}{2\sqrt{6}}$
- [ ] $\frac{\pi}{2\sqrt{3}}$
- [ ] none of the other choices

Q5

0 Points

Consider a probability distribution $\mathcal{P}(\mathbf{x}, y)$ that can be used to generate examples $(\mathbf{x}, y)$, and suppose we generate K i.i.d. examples from the distribution as validation examples, and store them in $\mathcal{D}_{\text{val}}$. For any fixed hypothesis h, we can show that

$$\text{Variance}_{\mathcal{D}_{\text{val}} \sim \mathcal{P}^K}\left[E_{\text{val}}(h)\right] = \square \cdot \text{Variance}_{(\mathbf{x},y)\sim\mathcal{P}}\left[\text{err}(h(\mathbf{x}), y)\right]$$

Which of the following is $\square$? Choose the correct answer; explain your answer.

- [ ] K
- [ ] $\sqrt{K}$
- [ ] $\frac{1}{\sqrt{K}}$
- [x] $\frac{1}{K}$
- [ ] none of the other choices

Q6

0 Points

Consider a binary classification algorithm $\mathcal{A}_{\text{majority}}$, which returns a constant classifier that always predicts the majority class (i.e., the class with more instances in the data set that it sees). As you can imagine, the returned classifier is the best-E_{in} one among all constant classifiers. For a binary classification data set with N positive examples and N negative examples, what is $E_{\text{loocv}}(\mathcal{A}_{\text{majority}})$? Choose the correct answer; explain your answer.

- [ ] $\frac{1}{N-1}$
- [ ] $\frac{1}{N}$
- [ ] $\frac{1}{N+1}$
- [x] 1
- [ ] none of the other choices

Q7

0 Points

Consider the decision stump model and the data generation process of generate x by a uniform distribution in $[-1, +1]$ and $y = \text{sign}(x)$. Use the generation process to generate a data set of N examples (instead of 2). If the data set contains at least two positive examples and at least two negative examples, which of the following is the tightest upper bound on the leave-one-out error of the decision stump model? Choose the correct answer; explain your answer.

- [ ] 0
- [ ] $\frac{1}{N}$
- [x] $\frac{2}{N}$
- [ ] $\frac{1}{2}$
- [ ] 1

Q8

0 Points

Consider N ''linearly separable'' 1D examples $\{(x_n, y_n)\}_{n=1}^{N}$. That is, $x_n \in \mathbb{R}$. Without loss of generality, assume that $x_1 \le x_2 \le \ldots x_M < x_{M+1} \le x_{M+2} \ldots \le x_N$, $y_n = -1$ for $n = 1, 2, \ldots, M$, and $y_n = +1$ for $n = M + 1, M + 2, \ldots, N$. Apply hard-margin SVM without transform on this data set. What is the largest margin achieved? Choose the correct answer; explain your answer.

- [ ] $\frac{1}{2}\left(x_N - x_M\right)$
- [ ] $\frac{1}{2}\left(x_{M+1} - x_1\right)$
- [ ] $\frac{1}{2}\left(\frac{1}{N-M}\sum_{n=M+1}^{N} x_n - \frac{1}{M}\sum_{n=1}^{M} x_n\right)$
- [ ] $\frac{1}{2}\left(x_N - x_1\right)$
- [x] $\frac{1}{2}\left(x_{M+1} - x_M\right)$

Q9

0 Points

In some situations, we expect to achieve a smaller margin for the positive examples and a larger margin for the negative examples. For instance, when there are very few negative examples and a lot more positive examples, giving the nagaive examples a smaller margin could be more robust. Consider an *uneven-margin* support vector machine that solves

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & (\mathbf{w}^T\mathbf{x_n} + b) \geq 1 \text{ for } y_n = +1 \\ & -(\mathbf{w}^T\mathbf{x_n} + b) \geq 1126 \text{ for } y_n = -1. \end{aligned}$$

Given the following examples.

$$\begin{aligned} \mathbf{x}_1 &= (0,4) \;\; y_1 = +1 \\ \mathbf{x}_2 &= (2,0) \;\; y_2 = -1 \\ \mathbf{x}_3 &= (-1,0) \;\; y_3 = +1 \\ \mathbf{x}_4 &= (0,0) \;\; y_4 = +1 \end{aligned}$$

What is the optimal $\mathbf{w}$ and b? Choose the correct answer; explain your answer.

- [ ] the optimal $\mathbf{w} = (\frac{-1127}{3}, 0), b = 1$
- [ ] the optimal $\mathbf{w} = (\frac{-1125}{2}, 0), b = -1$
- [ ] the optimal $\mathbf{w} = (\frac{-1125}{3}, 0), b = -1$
- [ ] the optimal $\mathbf{w} = (0, \frac{1127}{4}), b = 1$
- [x] the optimal $\mathbf{w} = (\frac{-1127}{2}, 0), b = 1$

Q10

0 Points

For a set of examples $\{(\mathbf{x_n}, y_n)\}_{n=1}^{N}$ and a kernel function K, consider a hypothesis set that contains

$$h_{\alpha,b}(\mathbf{x}) = \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n K(\mathbf{x_n}, \mathbf{x}) + b\right).$$

The classifier returned by SVM can be viewed as one such $h_{\alpha,b}$, where the values of α is determined by the dual QP solver and b is calculated from the KKT conditions.
In this problem, we study a simpler form of $h_{\alpha,b}$ where $h_\alpha = \mathbf{1}$ (the vector of all 1's) and $b = 0$. Let us name $h_{\mathbf{1},0}$ as $\hat{h}$ for simplicity. We will show that when using the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, if γ is large enough, $E_{in}(\hat{h}) = 0$. That is, when using the Gaussian kernel, we can ''easily'' separate the given data set if γ is large enough.
Assume that the distance between any pair of different $(\mathbf{x_n}, \mathbf{x_m})$ in the $\mathcal{X}$-space is no less than ϵ. That is,

$$\|\mathbf{x_n} - \mathbf{x_m}\| \geq \epsilon \quad \forall n \neq m.$$

What is the tightest lower bound of γ that ensures $E_{in}(\hat{h}) = 0$?
Choose the correct answer; explain your answer.

- [ ] $\frac{\ln^2(N+1)}{\epsilon^2}$
- [ ] $\frac{\ln(N+1)}{\epsilon^2}$
- [ ] $\frac{\ln(N)}{\epsilon^2}$
- [x] $\frac{\ln(N-1)}{\epsilon^2}$
- [ ] $\frac{\ln^2(N-1)}{\epsilon^2}$

Q11

0 Points

For any feature transform ϕ from $\mathcal{X}$ to $\mathcal{Z}$, the squared distance between two examples $\mathbf{x}$ and $\mathbf{x}'$ is $\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|^2$ in the $\mathcal{Z}$-space. For the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, compute the distance with the kernel trick. Then, for any two examples $\mathbf{x}$ and $\mathbf{x}'$, among the choices, what is the tightest upper bound for their distance in the $\mathcal{Z}$-space? Choose the correct answer; explain your answer.

- [ ] 0.0
- [ ] 0.5
- [ ] 1.0
- [x] 1.5
- [ ] 2.0

Q12

20 Points

Select the best λ^* *in a cheating manner* as $\arg\min_{\log_{10}\lambda \in \{-6,-3,0,3,6\}} E_{out}(\mathbf{w}_\lambda)$. Break the tie, if any, by selecting the largest λ.
What is $\log_{10}(\lambda^*)$? Choose the closest answer; provide your command/code.

- [ ] -6
- [ ] -3
- [ ] 0
- [x] 3
- [ ] 6

Q13
20 Points

Select the best λ^* as $\arg\min_{\log_{10}\lambda \in \{-6,-3,0,3,6\}} E_{in}(\mathbf{w}_\lambda)$.
Break the tie, if any, by selecting the largest λ.
What is $\log_{10}(\lambda^*)$? Choose the closest answer; provide your command/code.

- [ ] -6
- [ ] -3
- [x] 0
- [ ] 3
- [ ] 6

Q14
20 Points

Now randomly split the given training examples in $\mathcal{D}$ to two sets: 120 examples as $\mathcal{D}_{\text{train}}$ and 80 as $\mathcal{D}_{\text{val}}$. Run $\mathcal{A}_\lambda$ on *only* $\mathcal{D}_{\text{train}}$ to get $\mathbf{w}_\lambda^-$ (the weight vector within the g^- returned), and validate $\mathbf{w}_\lambda^-$ with $\mathcal{D}_{\text{val}}$ to get $E_{val}(\mathbf{w}_\lambda^-)$.
Select the best λ^* as

$$\underset{\log_{10}\lambda\in\{-6,-3,0,3,6\}}{\arg\min} \quad E_{val}(\mathbf{w}_\lambda^-).$$

Break the tie, if any, by selecting the largest λ. Repeat the experiment for 256 times, each with a different random split. What is the λ that is selected the most often? Choose the closest answer; provide your command/code.

- [ ] -6
- [ ] -3
- [ ] 0
- [x] 3
- [ ] 6

Q15
20 Points

Repeat the 256 experiments in the previous problem, and estimate $E_{out}(\mathbf{w}_{\lambda^*}^-)$ with the test set in each round of the experiments. What is the average value of $E_{out}(\mathbf{w}_{\lambda^*}^-)$? Choose the closest answer; provide your command/code.

- [ ] 0.13
- [ ] 0.15
- [x] 0.17
- [ ] 0.19
- [ ] 0.21

Q16

20 Points

Repeat the 256 experiments in the previous problem, but run $\mathcal{A}_\lambda$ on *the full* $\mathcal{D}$ to get $\mathbf{w}_\lambda$ instead.

Then, estimate $E_{out}(\mathbf{w}_{\lambda^*})$ with the test set. What is the average value of $E_{out}(\mathbf{w}_{\lambda^*})$? Choose the closest answer; provide your command/code.

- [ ] 0.13
- [x] 0.15
- [ ] 0.17
- [ ] 0.19
- [ ] 0.21

Q17
20 Points

Now randomly split the given training examples in $\mathcal{D}$ to five folds, the first 40 being fold 1, the next 40 being fold 2, and so on.
Select the best λ^* as

$$\underset{\log_{10}\lambda\in\{-6,-3,0,3,6\}}{\arg\min} \quad E_{cv}(\mathcal{A}_\lambda).$$

Break the tie, if any, by selecting the largest λ. Repeat the experiment for 256 times.
What is the average value of $E_{cv}(\mathcal{A}_{\lambda^*})$ Choose the closest answer; provide your command/code.

- [x] 0.13
- [ ] 0.15
- [ ] 0.17
- [ ] 0.19
- [ ] 0.21

Q18

20 Points

For L1-regularized logistic regression, select the best λ^* *in a cheating manner* as

$$\underset{\log_{10}\lambda\in\{-6,-3,0,3,6\}}{\arg\min} E_{out}(\mathbf{w}_\lambda).$$

Break the tie, if any, by selecting the largest λ.
What is $\log_{10}(\lambda^*)$? Choose the closest answer; provide your command/code.

- [ ] -6
- [ ] -3
- [x] 0
- [ ] 3
- [ ] 6

Q19

20 Points

Based on the λ^* chosen in the previous problem, obtain $\mathbf{w}_{\lambda^*}$ from L1-regularized logistic regression. How sparse is $\mathbf{w}_{\lambda^*}$? That is, how many components w_i within $\mathbf{w}_{\lambda^*}$ satisfies $|w_i| \leq 10^{-6}$? Choose the closest answer; provide your command/code.

- [ ] 1
- [ ] 200
- [ ] 400
- [ ] 800
- [x] 1000

Q20

20 Points

Based on the λ^* chosen in the Problem **12**, obtain $\mathbf{w}_{\lambda^*}$ from **L2-regularized** logistic regression. How sparse is $\mathbf{w}_{\lambda^*}$? That is, how many components w_i within $\mathbf{w}_{\lambda^*}$ satisfies $|w_i| \leq 10^{-6}$? Choose the closest answer; provide your command/code.

- [x] 1
- [ ] 200
- [ ] 400
- [ ] 800
- [ ] 1000

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