作業四

TOTAL POINTS 200

1. **Deterministic noise depends on $\mathcal{H}$, as some models approximate f better than others. Assume $\mathcal{H}' \subset \mathcal{H}$ and that f is fixed. In general (but not necessarily in all cases), if we use $\mathcal{H}'$ instead of $\mathcal{H}$, how does deterministic noise behave?**

10 points

○ In general, deterministic noise will decrease.

○ In general, deterministic noise will increase.

○ In general, deterministic noise will be the same.

○ If $d_{\text{vc}}(\mathcal{H}') \leq \frac{1}{2}\, d_{\text{vc}}(\mathcal{H})$, deterministic noise will increase, else it will decrease.

○ If $d_{\text{vc}}(\mathcal{H}') \leq \frac{1}{2}\, d_{\text{vc}}(\mathcal{H})$, deterministic noise will decrease, else it will increase.

2. **Consider the following hypothesis set for $\mathbf{x} \in \mathbb{R}^d$ defined by the constraint:**

10 points

$$\mathcal{H}(d, d_0) = \{h \mid h(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}; w_i = 0 \text{ for } i \geq d_0\},$$

which of the following statements is correct?

○ $\mathcal{H}(10,3) \subset \mathcal{H}(10,4)$

○ $\mathcal{H}(10,3) \cup \mathcal{H}(10,4) = \{\}$

○ $\mathcal{H}(10,3) \supset \mathcal{H}(10,4)$

○ $\mathcal{H}(10,3) \cap \mathcal{H}(10,4) = \{\}$

○ none of the other choices

3. **For Questions 3-4, consider the augmented error $E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\, \mathbf{w}^T\mathbf{w}$with some $\lambda > 0$. If we want to minimize the augmented error $E_{\text{aug}}(\mathbf{w})$ by gradient descent with η as learning rate, which of the following is a correct update rule?**

10 points

○ $\mathbf{w}(t+1) \longleftarrow \mathbf{w}(t) + \eta\lambda\nabla E_{\text{in}}(\mathbf{w}(t)).$

○ $\mathbf{w}(t+1) \longleftarrow \mathbf{w}(t) - \eta\lambda\nabla E_{\text{in}}(\mathbf{w}(t)).$

○ $\mathbf{w}(t+1) \longleftarrow (1 - \frac{2\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t)).$

○ $\mathbf{w}(t+1) \longleftarrow (1 + \frac{2\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t)).$

○ none of the other choices

4. **Let $\mathbf{w}_{\text{lin}}$ be the optimal solution for the plain-vanilla linear regression and $\mathbf{w}_{\text{reg}}(\lambda)$ be the optimal solution for minimizing E_{aug} in Question 3, with E_{in} being the squared error for linear regression. Which of the following is correct?**

10 points

○ $\|\mathbf{w}_{\text{reg}}(\lambda)\|$ is always a non-decreasing function of λ for $\lambda \geq 0$

○ $\|\mathbf{w}_{\text{reg}}(\lambda)\| \leq \|\mathbf{w}_{\text{lin}}\|$ for any $\lambda > 0$

○ none of the other choices

○ $\|\mathbf{w}_{\text{reg}}(\lambda)\|$ is always a constant function of λ for $\lambda \geq 0$

○ $\|\mathbf{w}_{\text{reg}}(\lambda)\| \geq \|\mathbf{w}_{\text{lin}}\|$ for any $\lambda > 0$

5. **You are given the data points: $(-1, 0), (\varrho, 1), (1, 0),\ \varrho \geq 0$, and a choice between two models:**

10 points

- **constant $h_0(x) = b_0$ and**
- **linear $h_1(x) = a_1x + b_1$.**

For which value of ϱ would the two models be tied using leave-one-out cross-validation with the squared error measure?

○ $\sqrt{\sqrt{3}+4}$

○ $\sqrt{\sqrt{3}-1}$

○ $\sqrt{9+4\sqrt{6}}$

○ $\sqrt{9-\sqrt{6}}$

○ none of the other choices

6. **For Questions 6-7, suppose that for 5 weeks in a row, a letter arrives in the mail that predicts the outcome of the upcoming Monday night baseball game. Assume there are no tie. You keenly watch each Monday and to your surprise, the prediction is correct each time. On the day after the fifth game, a letter arrives, stating that if you wish to see next week's prediction, a payment of NTD 1000 is required. Which of the following statement is true?**

10 points

○ There are 31 win-lose predictions for 5 games.

○ If the sender wants to make sure that at least one person receives correct predictions on all 5 games from him, the sender should target to begin with at least 5 people.

○ To make sure that at least one person receives correct predictions on all $5$games from the sender, after the first letter `predicts' the outcome of the first game, the sender should target at least 16 people with the second letter.

○ To make sure that at least one person receives correct predictions on all $5$games from him, at least 64 letters should be sent before the fifth game.

○ none of the other choices

7. **If the cost of printing and mailing out each letter is NTD 10. If the sender sends the minimum number of letters out, how much money can be made for the above `fraud' to succeed once? That is, one of the recipients does send him NTD 1000 to receive the prediction of the 6-th game?**

10 points

○ NTD 340

○ NTD 370

○ NTD 400

○ NTD 430

○ NTD 460

8. **For Questions 8-10, please read the following story first. In our credit card example, the bank starts with some vague idea of what constitutes a good credit risk. So, as customers $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ arrive, the bank applies its vague idea to approve credit cards for some of these customers based on a formula $a(\mathbf{x})$. Then, only those who get credit cards are monitored to see if they default or not.**

10 points

For simplicity, suppose that the first $N = 10000$ customers were given credit cards by the credit approval function $a(\mathbf{x})$. Now that the bank knows the behavior of these customers, it comes to you to improve their algorithm for approving credit. The bank gives you the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. Before you look at the data, you do mathematical derivations and come up with a credit approval function. You now test it on the data and, to your delight, obtain perfect prediction.

What is M, the size of your hypothesis set?

○ 1

○ N

○ 2^N

○ N^2

○ We have no idea about it.

9. **With such an M, what does the Hoeffding bound say about the probability that the true average error rate of g is worse than 1% for $N = 10,000$?** 10 points

○ ≤ 0.171

○ ≤ 0.221

○ ≤ 0.271

○ ≤ 0.321

○ none of the other choices

10. **You assure the bank that you have a got a system g for approving credit cards for new customers, which is nearly error-free. Your confidence is given by your answer to the previous question. The bank is thrilled and uses your g to approve credit for new customers. To their dismay, more than half their credit cards are being defaulted on. Assume that the customers that were sent to the old credit approval function and the customers that were sent to your g are indeed i.i.d. from the same distribution, and the bank is lucky enough (so the "bad luck" that "the true error of g is worse than 1%" does not happen). Which of the following claim is true?** 10 points

○ By applying $a(\mathbf{x})$ NOR $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

○ By applying $a(\mathbf{x})$ NAND $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

○ By applying $a(\mathbf{x})$ OR $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

○ By applying $a(\mathbf{x})$ AND $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

○ none of the other choices

11. **For Questions 11-12, consider linear regression with virtual examples. That is, we** 10 points

add K virtual examples $(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_K, \tilde{y}_K)$ to the training data set, and solve

$$\min_{\mathbf{w}} \frac{1}{N+K} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \sum_{k=1}^{K} (\tilde{y}_k - \mathbf{w}^T \tilde{\mathbf{x}}_k)^2 \right).$$

We will show that using some "special" virtual examples, which were claimed to be a possible way to combat overfitting in Lecture 9, is related to regularization, another possible way to combat overfitting discussed in Lecture 10. Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1 \tilde{\mathbf{x}}_2 \ldots \tilde{\mathbf{x}}_K]^T$, and $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K]^T$.

What is the optimal $\mathbf{w}$ to the optimization problem above, assuming that all the inversions exist?

○ $(\mathbf{X}^T\mathbf{X})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$

○ $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$

○ $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$

○ $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$

○ none of the other choices

12. For what $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ will the solution of the linear regression problem above equal to

10 points

$$\mathbf{w}_{\text{reg}} = \text{argmin}_{\mathbf{w}} \frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2?$$

○ $\tilde{\mathbf{X}} = \mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$

○ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$

○ $\tilde{\mathbf{X}} = \lambda\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{1}$

○ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{X}, \tilde{\mathbf{y}} = \mathbf{y}$

○ none of the other choices

13. Consider regularized linear regression (also called ridge regression) for classification

10 points

$$\mathbf{w}_{\text{reg}} = \text{argmin}_{\mathbf{w}} \left(\frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right).$$

Run the algorithm on the following data set as $\mathcal{D}$:

https://www.csie.ntu.edu.tw/~htlin/mooc/datasets/mlfound_algo/hw4_train.dat

and the following set for evaluating E_{out}:

https://www.csie.ntu.edu.tw/~htlin/mooc/datasets/mlfound_algo/hw4_test.dat

Because the data sets are for classification, please consider only the 0/1 error for all Questions below.

Let $\lambda = 10$, which of the followings is the corresponding E_{in} and E_{out}?

○ $E_{in} = 0.015, E_{out} = 0.020$

○ $E_{in} = 0.030, E_{out} = 0.015$

○ $E_{in} = 0.035, E_{out} = 0.020$

○ $E_{in} = 0.050, E_{out} = 0.045$

○ $E_{in} = 0.020, E_{out} = 0.010$

14. Following the previous Question, aong $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{in}? Compute λ and its corresponding E_{in} and E_{out} then select the closest answer. Break the tie by selecting the largest λ.

10 points

○ $\log_{10} \lambda = -2, E_{in} = 0.030, E_{out} = 0.040$

○ $\log_{10} \lambda = -4, E_{in} = 0.015, E_{out} = 0.020$

○ $\log_{10} \lambda = -6, E_{in} = 0.030, E_{out} = 0.040$

○ $\log_{10} \lambda = -8, E_{in} = 0.015, E_{out} = 0.020$

○ $\log_{10} \lambda = -10, E_{in} = 0.030, E_{out} = 0.040$

15. Following the previous Question, among $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{out}? Compute λ and the corresponding E_{in} and E_{out} then select the closest answer. Break the tie by selecting the largest λ.

10 points

○ $\log_{10} \lambda = -1, E_{in} = 0.015, E_{out} = 0.015$

○ $\log_{10}\lambda = -3, E_{in} = 0.015, E_{out} = 0.015$

○ $\log_{10}\lambda = -5, E_{in} = 0.015, E_{out} = 0.030$

○ $\log_{10}\lambda = -7, E_{in} = 0.030, E_{out} = 0.015$

○ $\log_{10}\lambda = -9, E_{in} = 0.030, E_{out} = 0.030$

16. Now split the given training examples in $\mathcal{D}$ to the first 120 examples for $\mathcal{D}_{\text{train}}$ and 80 for $\mathcal{D}_{\text{val}}$. \textit{Ideally, you should randomly do the $120/80$ split. Because the given examples are already randomly permuted, however, we would use a fixed split for the purpose of this problem.}

10 points

Run the algorithm on $\mathcal{D}_{\text{train}}$ to get g^{-}_{λ}, and validate g^{-}_{λ} with $\mathcal{D}_{\text{val}}$. Among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum $E_{train}(g^{-}_{\lambda})$? Compute λ and the corresponding $E_{train}(g^{-}_{\lambda})$, $E_{val}(g^{-}_{\lambda})$ and $E_{out}(g^{-}_{\lambda})$ then select the closet answer. Break the tie by selecting the largest λ.

○ $\log_{10}\lambda = 0, E_{train}(g^{-}_{\lambda}) = 0.000, E_{val}(g^{-}_{\lambda}) = 0.050, E_{out}(g^{-}_{\lambda}) = 0.025$

○ $\log_{10}\lambda = -2, E_{train}(g^{-}_{\lambda}) = 0.010, E_{val}(g^{-}_{\lambda}) = 0.050, E_{out}(g^{-}_{\lambda}) = 0.035$

○ $\log_{10}\lambda = -4, E_{train}(g^{-}_{\lambda}) = 0.000, E_{val}(g^{-}_{\lambda}) = 0.010, E_{out}(g^{-}_{\lambda}) = 0.035$

○ $\log_{10}\lambda = -6, E_{train}(g^{-}_{\lambda}) = 0.010, E_{val}(g^{-}_{\lambda}) = 0.010, E_{out}(g^{-}_{\lambda}) = 0.025$

○ $\log_{10}\lambda = -8, E_{train}(g^{-}_{\lambda}) = 0.000, E_{val}(g^{-}_{\lambda}) = 0.050, E_{out}(g^{-}_{\lambda}) = 0.025$

17. Following the previous Question, among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum $E_{val}(g^{-}_{\lambda})$? Compute λ and the corresponding $E_{train}(g^{-}_{\lambda})$, $E_{val}(g^{-}_{\lambda})$ and $E_{out}(g^{-}_{\lambda})$then select the closet answer. Break the tie by selecting the largest λ.

10 points

○ $\log_{10}\lambda = 0, E_{train}(g^{-}_{\lambda}) = 0.033, E_{val}(g^{-}_{\lambda}) = 0.038, E_{out}(g^{-}_{\lambda}) = 0.028$

○ $\log_{10}\lambda = -3, E_{train}(g^{-}_{\lambda}) = 0.000, E_{val}(g^{-}_{\lambda}) = 0.028, E_{out}(g^{-}_{\lambda}) = 0.038$

○ $\log_{10}\lambda = -6, E_{train}(g^{-}_{\lambda}) = 0.066, E_{val}(g^{-}_{\lambda}) = 0.038, E_{out}(g^{-}_{\lambda}) = 0.038$

○ $\log_{10}\lambda = -9, E_{train}(g^{-}_{\lambda}) = 0.033, E_{val}(g^{-}_{\lambda}) = 0.028, E_{out}(g^{-}_{\lambda}) = 0.028$

○ $\log_{10}\lambda = -10, E_{train}(g^{-}_{\lambda}) = 0.066, E_{val}(g^{-}_{\lambda}) = 0.028, E_{out}(g^{-}_{\lambda}) = 0.028$

18. Run the algorithm with the optimal λ of the previous Question on the whole $\mathcal{D}$to get g_{λ}. Compute $E_{in}(g_{\lambda})$ and $E_{out}(g_{\lambda})$ then select the closet answer.

10 points

○ $E_{in}(g_\lambda) = 0.015, E_{out}(g_\lambda) = 0.020$

○ $E_{in}(g_\lambda) = 0.025, E_{out}(g_\lambda) = 0.030$

○ $E_{in}(g_\lambda) = 0.035, E_{out}(g_\lambda) = 0.020$

○ $E_{in}(g_\lambda) = 0.045, E_{out}(g_\lambda) = 0.030$

○ $E_{in}(g_\lambda) = 0.055, E_{out}(g_\lambda) = 0.020$

19. For Questions 19-20, 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. Again, we take a fixed split because the given examples are already randomly permuted.

10 points

Among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{cv}, where E_{cv} comes from the five folds defined above? Compute $\lambda$and the corresponding E_{cv} then select the closet answer. Break the tie by selecting the largest λ.

○ $\log_{10}\lambda = 0, E_{cv} = 0.030$

○ $\log_{10}\lambda = -2, E_{cv} = 0.020$

○ $\log_{10}\lambda = -4, E_{cv} = 0.030$

○ $\log_{10}\lambda = -6, E_{cv} = 0.020$

○ $\log_{10}\lambda = -8, E_{cv} = 0.030$

20. Run the algorithm with the optimal λ of the previous problem on the whole $\mathcal{D}$to get g_λ. Compute $E_{in}(g_\lambda)$ and $E_{out}(g_\lambda)$ then select the closet answer.

10 points

○ $E_{in}(g_\lambda) = 0.005, E_{out}(g_\lambda) = 0.010$

○ $E_{in}(g_\lambda) = 0.015, E_{out}(g_\lambda) = 0.020$

○ $E_{in}(g_\lambda) = 0.025, E_{out}(g_\lambda) = 0.020$

○ $E_{in}(g_\lambda) = 0.035, E_{out}(g_\lambda) = 0.030$

○ $E_{in}(g_\lambda) = 0.045, E_{out}(g_\lambda) = 0.020$