Feedback — 作業四

Help Center

You submitted this quiz on **Wed 11 Nov 2015 7:59 AM CST**. You got a score of **375.00** out of **400.00**. You can attempt again, if you'd like.

Question 1

Overfitting and Deterministic Noise

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?

Your Answer		Score	Explanation
○ In general, deterministic noise will decrease.			
◉ In general, deterministic noise will increase.	✔	20.00	
○ In general, deterministic noise will be the same.			
○ If $d_{\text{vc}}(\mathcal{H}') \le \frac{1}{2} d_{\text{vc}}(\mathcal{H})$, deterministic noise will increase, else it will decrease.			
○ If $d_{\text{vc}}(\mathcal{H}') \le \frac{1}{2} d_{\text{vc}}(\mathcal{H})$, deterministic noise will decrease, else it will increase.			
Total		20.00 / 20.00	

Question 2

Regularization With Polynomials Polynomial models can be viewed as linear models in a space $\mathcal{Z}$, under a nonlinear transform $\Phi : \mathcal{X} \to \mathcal{Z}$. Here, Φ transforms the scalar x into a vector $\mathbf{z}$ of Legendre polynomials, $\mathbf{z} = (1, L_1(x), L_2(x), \ldots, L_Q(x))$. Our hypothesis set will be expressed as a linear combination of these polynomials,
$\mathcal{H}_Q = \left\{ h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} = \sum_{q=0}^{Q} w_q L_q(x) \right\}$, where $L_0(x) = 1$.

Consider the following hypothesis set defined by the constraint:

$\mathcal{H}(Q, c, Q_o) = \{h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} \in \mathcal{H}_Q; w_q = c \text{ for } q \ge Q_o\}$, which of the following statements is correct?

Your Answer		Score	Explanation
$\mathcal{H}(10,1,3) \cap \mathcal{H}(10,1,4) = \mathcal{H}_1$			
$\mathcal{H}(10,0,3) \cup \mathcal{H}(10,0,4) = \mathcal{H}_4$			
$\mathcal{H}(10,0,3) \cap \mathcal{H}(10,0,4) = \mathcal{H}_2$	✔	20.00	
$\mathcal{H}(10,0,3) \cup \mathcal{H}(10,1,4) = \mathcal{H}_3$			
none of the other choices			
Total		20.00 / 20.00	

Question 3

Regularization and Weight Decay 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 followings are the correct update rules?

Your Answer		Score	Explanation
$\mathbf{w}(t+1) \longleftarrow (1 - \frac{2\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t))$.	✔	20.00	
$\mathbf{w}(t+1) \longleftarrow \mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t))$.			
$\mathbf{w}(t+1) \longleftarrow (1 - \frac{\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t))$.			
$\mathbf{w}(t+1) \longleftarrow (1 + \frac{\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t))$.			
$\mathbf{w}(t+1) \longleftarrow (1 + \frac{2\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t))$.			
Total		20.00 / 20.00	

Question 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 the formulation above. Select all the correct statements below:

Your Answer		Score	Explanation
none of the other choices			
$\|\mathbf{w}_{\text{reg}}(\lambda)\|$ is a non-decreasing function of λ for $\lambda \geq 0$			
$\|\mathbf{w}_{\text{reg}}(\lambda)\|$ is a non-increasing function of λ for $\lambda \geq 0$	✔	20.00	
$\|\mathbf{w}_{\text{reg}}(\lambda)\| > \|\mathbf{w}_{\text{lin}}\|$ for any $\lambda > 0$			
$\|\mathbf{w}_{\text{reg}}(\lambda)\| = \|\mathbf{w}_{\text{lin}}\|$ for any $\lambda > 0$			
Total		20.00 / 20.00	

Question 5

Leave-One-Out Cross-Validation

You are given the data points: $(-1, 0), (\rho, 1), (1, 0), \ \rho \geq 0$, and a choice between two models: constant [$h_0(x) = b_0$] and linear [$h_1(x) = a_1 x + b_1$]. For which value of ρ would the two models be tied using leave-one-out cross-validation with the squared error measure?

Your Answer		Score	Explanation
$\sqrt{\sqrt{3}+4}$			
$\sqrt{\sqrt{3}-1}$			
$\sqrt{9+4\sqrt{6}}$	✔	20.00	
$\sqrt{9-\sqrt{6}}$			
none of the other choices			
Total		20.00 / 20.00	

Question 6

Learning Principles In Problems 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 1,000 is required.

Which of the following statement is true?

Your Answer		Score	Explanation
◉ To make sure that at least one person receives correct predictions on all 5 games from him, after the first letter `predicts' the outcome of the first game, the sender should target at least 16 people with the second letter.	✔	20.00	
○ none of the other choices			
○ To make sure that at least one person receives correct predictions on all 5 games from him, at least 64 letters are sent before the fifth game.			
○ 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.			
○ none of the other choices			
Total		20.00 / 20.00	

Question 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 1,000 to receive the prediction of the 6-th game?

Your Answer		Score	Explanation
◉ NTD 370	✔	20.00	
○ NTD 340			
○ NTD 430			
○ NTD 460			
○ NTD 400			
Total		20.00 / 20.00	

Question 8

For Problems 8-10, we consider the following. 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. For simplicity, suppose that the first $N = 10,000$ 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?

Your Answer		Score	Explanation
◉ 2^N	✖	-5.00	
○ N			
○ 1			
○ N^2			
○ We have no idea about it.			
Total		-5.00 / 20.00	

Question 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 $N = 10,000$?

Your Answer		Score	Explanation
○ ≤ 0.171			
○ ≤ 0.221			
◉ ≤ 0.271	✔	20.00	
○ ≤ 0.321			

≤ 0.371

Total	20.00 / 20.00

Question 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). Select all the true claims for this situation.

Your Answer		Score	Explanation
By applying $a(\mathbf{x})$ XNOR $g(\mathbf{x})$ 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			
By applying $a(\mathbf{x})$ XOR $g(\mathbf{x})$ 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.	✔	20.00	
Total		20.00 / 20.00	

Question 11

Virtual Examples and Regularization

Consider linear regression with virtual examples. That is, we 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?

Your Answer		Score	Explanation
○ none of the other choices			
○ $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$			
◉ $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$	✔	20.00	
○ $(\mathbf{X}^T\mathbf{X})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$			
○ $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$			
Total		20.00 / 20.00	

Question 12

For what $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ will the solution of this linear regression equal to

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

Your Answer	Score	Explanation
○ $\tilde{\mathbf{X}} = \lambda\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{1}$		
○ none of the other choices		
○ $\tilde{\mathbf{X}} = \mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$		

○ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{X}, \tilde{\mathbf{y}} = \mathbf{y}$		
◉ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$	✔	20.00
Total		20.00 / 20.00

Question 13

Experiment with Regularized Linear Regression and Validation

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

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

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

https://d396qusza40orc.cloudfront.net/ntumlone%2Fhw4%2Fhw4_train.dat

and the following set for evaluating E_{out}

https://d396qusza40orc.cloudfront.net/ntumlone%2Fhw4%2Fhw4_test.dat

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

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

Your Answer		Score	Explanation
◉ $E_{in} = 0.050, E_{out} = 0.045$	✔	20.00	
○ $E_{in} = 0.030, E_{out} = 0.015$			
○ $E_{in} = 0.015, E_{out} = 0.020$			
○ $E_{in} = 0.020, E_{out} = 0.010$			
○ $E_{in} = 0.035, E_{out} = 0.020$			
Total		20.00 / 20.00	

Question 14

Among $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. 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.

Your Answer		Score	Explanation
◉ $\log_{10} \lambda = -8, E_{in} = 0.015, E_{out} = 0.020$	✔	20.00	
○ $\log_{10} \lambda = -10, E_{in} = 0.030, E_{out} = 0.040$			
○ $\log_{10} \lambda = -6, E_{in} = 0.030, E_{out} = 0.040$			
○ $\log_{10} \lambda = -2, E_{in} = 0.030, E_{out} = 0.040$			
○ $\log_{10} \lambda = -4, E_{in} = 0.015, E_{out} = 0.020$			
Total		20.00 / 20.00	

Question 15

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.

Your Answer		Score	Explanation
○ $\log_{10} \lambda = -3, E_{in} = 0.015, E_{out} = 0.015$			
○ $\log_{10} \lambda = -5, E_{in} = 0.015, E_{out} = 0.030$			
○ $\log_{10} \lambda = -9, E_{in} = 0.030, E_{out} = 0.030$			
○ $\log_{10} \lambda = -1, E_{in} = 0.015, E_{out} = 0.015$			
◉ $\log_{10} \lambda = -7, E_{in} = 0.030, E_{out} = 0.015$	✔	20.00	
Total		20.00 / 20.00	

Question 16

Now split the given training examples in $\mathcal{D}$ to the first 120 examples for $\mathcal{D}_{train}$ and 80 for $\mathcal{D}_{val}$.

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.

Run the algorithm on $\mathcal{D}_{train}$ to get g_λ^-, and validate g_λ^- with $\mathcal{D}_{val}$.
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.

Your Answer		Score	Explanation
○ $\log_{10}\lambda = -6, E_{train}(g_\lambda^-) = 0.010, E_{val}(g_\lambda^-) = 0.010, E_{out}(g_\lambda^-) = 0.025$			
◉ $\log_{10}\lambda = -8, E_{train}(g_\lambda^-) = 0.000, E_{val}(g_\lambda^-) = 0.050, E_{out}(g_\lambda^-) = 0.025$	✔	20.00	
○ $\log_{10}\lambda = -4, E_{train}(g_\lambda^-) = 0.000, E_{val}(g_\lambda^-) = 0.010, E_{out}(g_\lambda^-) = 0.035$			
○ $\log_{10}\lambda = -0, E_{train}(g_\lambda^-) = 0.000, E_{val}(g_\lambda^-) = 0.050, E_{out}(g_\lambda^-) = 0.025$			
○ $\log_{10}\lambda = -2, E_{train}(g_\lambda^-) = 0.010, E_{val}(g_\lambda^-) = 0.050, E_{out}(g_\lambda^-) = 0.035$			
Total		20.00 / 20.00	

Question 17

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.

Your Answer		Score	Explanation
○ $\log_{10}\lambda = -6, E_{train}(g_\lambda^-) = 0.066, E_{val}(g_\lambda^-) = 0.038, E_{out}(g_\lambda^-) = 0.038$			
◉ $\log_{10}\lambda = 0, E_{train}(g_\lambda^-) = 0.033, E_{val}(g_\lambda^-) = 0.038, E_{out}(g_\lambda^-) = 0.028$	✔	20.00	
○ $\log_{10}\lambda = -3, E_{train}(g_\lambda^-) = 0.000, E_{val}(g_\lambda^-) = 0.028, E_{out}(g_\lambda^-) = 0.038$			

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

$\log_{10}\lambda = -8, E_{train}(g_{\lambda}^{-}) = 0.066, E_{val}(g_{\lambda}^{-}) = 0.028, E_{out}(g_{\lambda}^{-}) = 0.028$

Total 20.00 / 20.00

Question 18

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.

Your Answer		Score	Explanation
$E_{in}(g_\lambda) = 0.025, E_{out}(g_\lambda) = 0.030$			
$E_{in}(g_\lambda) = 0.035, E_{out}(g_\lambda) = 0.020$	✔	20.00	
$E_{in}(g_\lambda) = 0.055, E_{out}(g_\lambda) = 0.020$			
$E_{in}(g_\lambda) = 0.045, E_{out}(g_\lambda) = 0.030$			
$E_{in}(g_\lambda) = 0.015, E_{out}(g_\lambda) = 0.020$			
Total		20.00 / 20.00	

Question 19

Now 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.

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 λ and the corresponding E_{cv} then select the closet answer. Break the tie by selecting the largest λ.

Your Answer	Score	Explanation

○ $\log_{10} \lambda = -2, E_{cv} = 0.020$		
○ $\log_{10} \lambda = -4, E_{cv} = 0.030$		
○ $\log_{10} \lambda = -6, E_{cv} = 0.020$		
○ $\log_{10} \lambda = 0, E_{cv} = 0.030$		
◉ $\log_{10} \lambda = -8, E_{cv} = 0.030$	✔	20.00
Total		20.00 / 20.00

Question 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.

Your Answer		Score	Explanation
◉ $E_{in}(g_\lambda) = 0.015, E_{out}(g_\lambda) = 0.020$	✔	20.00	
○ $E_{in}(g_\lambda) = 0.005, E_{out}(g_\lambda) = 0.010$			
○ $E_{in}(g_\lambda) = 0.045, E_{out}(g_\lambda) = 0.020$			
○ $E_{in}(g_\lambda) = 0.035, E_{out}(g_\lambda) = 0.030$			
○ $E_{in}(g_\lambda) = 0.025, E_{out}(g_\lambda) = 0.020$			
Total		20.00 / 20.00	