Feedback — 作業三

Help Center

You submitted this quiz on **Tue 10 Nov 2015 11:25 PM PST**. You got a score of **400.00** out of **400.00**. However, you will not get credit for it, since it was submitted past the deadline.

Question 1

Questions 1-2 are about linear regression.

Consider a noisy target $y = \mathbf{w}_f^T \mathbf{x} + \epsilon$, where $\mathbf{x} \in \mathbb{R}^d$ (with the added coordinate $x_0 = 1$), $y \in \mathbb{R}$, $\mathbf{w}_f$ is an unknown vector, and ϵ is a noise term with zero mean and σ^2 variance. Assume ϵ is independent of $\mathbf{x}$ and of all other ϵ's. If linear regression is carried out using a training data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, and outputs the parameter vector $\mathbf{w}_{\text{lin}}$, it can be shown that the expected in-sample error E_{in} with respect to $\mathcal{D}$ is given by:

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$$

For $\sigma = 0.1$ and $d = 8$, which among the following choices is the smallest number of examples N that will result in an expected E_{in} greater than 0.008?

Your Answer		Score	Explanation
◉ 100	✔	20.00	
○ 1000			
○ 500			
○ 25			
○ 10			
Total		20.00 / 20.00	

Question 2

Recall that we have introduced the hat matrix $\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$ in class, where $\mathrm{X} \in \mathbb{R}^{N\times(d+1)}$ for N examples and d features. Assume $\mathrm{X}^T\mathrm{X}$ is invertible, which statements of H are true? (a) H is always positive semi-definite. (b) H is always invertible. (c) some eigenvalues of H are possibly bigger than 1. (d) $d+1$ eigenvalues of H are exactly 1. (e) $\mathrm{H}^{1126} = \mathrm{H}$.

Your Answer		Score	Explanation
◉ none of the other choices	✔	20.00	
○ ace			
○ cd			
○ c			
○ abcde			
Total		20.00 / 20.00	

Question 3

Questions 3-5 are about error and SGD

Which of the following is an upper bound of $\left[\mathrm{sign}(\mathbf{w}^T\mathbf{x}) \neq y\right]$ for $y \in \{-1, +1\}$?

Your Answer		Score	Explanation
○ none of the other choices			
○ $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$			
◉ $err(\mathbf{w}) = (\max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2$	✔	20.00	
○ $err(\mathbf{w}) = \theta(-y\mathbf{w}^T\mathbf{x})$			
○ $err(\mathbf{w}) = (-y\mathbf{w}^T\mathbf{x})$			
Total		20.00 / 20.00	

Question 4

Which of the following is not a everywhere-differentiable function of $\mathbf{w}$?

Your Answer		Score	Explanation
$err(\mathbf{w}) = (-y\mathbf{w}^T\mathbf{x})$			
$err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$	✔	20.00	
$err(\mathbf{w}) = (\max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2$			
none of the other choices			
$err(\mathbf{w}) = \theta(-y\mathbf{w}^T\mathbf{x})$			
Total		20.00 / 20.00	

Question 5

When using SGD on the following error functions and `ignoring' some singular points that are not differentiable, which of the following error function results in PLA?

Your Answer		Score	Explanation
none of the other choices			
$err(\mathbf{w}) = \left(\max(0, 1 - y\mathbf{w}^T\mathbf{x})\right)^2$			
$err(\mathbf{w}) = (-y\mathbf{w}^T\mathbf{x})$			
$err(\mathbf{w}) = \theta(-y\mathbf{w}^T\mathbf{x})$			
$err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$	✔	20.00	
Total		20.00 / 20.00	

Question 6

For Questions 6-10, you will play with gradient descent algorithm and variants. Consider a function

$$E(u,v) = e^u + e^{2v} + e^{uv} + u^2 - 2uv + 2v^2 - 3u - 2v.$$

What is the gradient $\nabla E(u,v)$ around $(u,v) = (0,0)$?

Your Answer		Score	Explanation
none of the other choices			
$(3,-1)$			
$(-2,0)$	✔	20.00	
$(-3,1)$			
$(0,-2)$			
Total		20.00 / 20.00	

Question 7

In class, we have taught that the update rule of the gradient descent algorithm is

$$(u_{t+1}, v_{t+1}) = (u_t, v_t) - \eta \nabla E(u_t, v_t)$$

Please start from $(u_0, v_0) = (0,0)$, and fix $\eta = 0.01$, what is $E(u_5, v_5)$ after five updates?

Your Answer		Score	Explanation
4.904			
3.277			
0.365			
1.436			
2.825	✔	20.00	

Total	20.00 / 20.00

Question 8

Continue from Question 7, if we approximate the $E(u+\Delta u, v+\Delta v)$ by $\hat{E}_2(\Delta u, \Delta v)$, where $\hat{E}_2$ is the second-order Taylor's expansion of E around (u, v). Suppose

$$\hat{E}_2(\Delta u, \Delta v) = b_{uu}(\Delta u)^2 + b_{vv}(\Delta v)^2 + b_{uv}(\Delta u)(\Delta v) + b_u \Delta u + b_v \Delta v + b.$$

What are the values of $(b_{uu}, b_{vv}, b_{uv}, b_u, b_v, b)$ around $(u, v) = (0, 0)$

Your Answer		Score	Explanation
○ $(1.5, 4, -0.5, -1, -2, 0)$			
○ none of the other choices			
○ $(3, 8, -1, -2, 0, 3)$			
○ $(3, 8, -0.5, -1, -2, 0)$			
◉ $(1.5, 4, -1, -2, 0, 3)$	✔	20.00	
Total		20.00 / 20.00	

Question 9

Continue from Question 8 and denote the Hessian matrix to be $\nabla^2 E(u, v)$, and assume that the Hessian matrix is positive definite. What is the optimal $(\Delta u, \Delta v)$ to minimize $\hat{E}_2(\Delta u, \Delta v)$? The direction is called the *Newton Direction*.

Your Answer		Score	Explanation
◉ $-\left(\nabla^2 E(u, v)\right)^{-1} \nabla E(u, v)$	✔	20.00	
○ $+\left(\nabla^2 E(u, v)\right)^{-1} \nabla E(u, v)$			
○ $-\nabla^2 E(u, v) \nabla E(u, v)$			

$+\nabla^2 E(u,v)\nabla E(u,v)$	
none of the other choices	
Total	20.00 / 20.00

Question 10

Using the Newton direction (without η) to update, please start from $(u_0, v_0) = (0, 0)$, what is $E(u_5, v_5)$ after five updates?

Your Answer		Score	Explanation
2.361	✔	20.00	
4.532			
1.279			
3.046			
0.356			
Total		20.00 / 20.00	

Question 11

For Questions 11-12, you will play with feature transforms

Consider six inputs $\mathbf{x}_1 = (1, 1), \mathbf{x}_2 = (1, -1), \mathbf{x}_3 = (-1, -1), \mathbf{x}_4 = (-1, 1), \mathbf{x}_5 = (0, 0), \mathbf{x}_6 = (1, 0)$. What is the biggest subset of those input vectors that can be shattered by the union of quadratic, linear, or constant hypotheses of $\mathbf{x}$?

Your Answer		Score	Explanation
$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$	✔	20.00	
$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$			
$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$			

$\mathbf{x}_1, \mathbf{x}_3$	
$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$	
Total	20.00 / 20.00

Question 12

Assume that a transformer peeks the data and decides the following transform $\mathbf{\Phi}$ ``intelligently'' from the data of size N. The transform maps $\mathbf{x} \in \mathbb{R}^d$ to $\mathbf{z} \in \mathbb{R}^N$, where

$$(\mathbf{\Phi}(\mathbf{x}))_n = z_n = [\mathbf{x} = \mathbf{x}_n]$$

Consider a learning algorithm that performs linear classification after the feature transform. That is, the algorithm effectively works on an $\mathcal{H}_{\mathbf{\Phi}}$ that includes all possible $\mathbf{\Phi}$. What is $d_{vc}(\mathcal{H}_{\mathbf{\Phi}})$ (i.e. the maximum number of points that can be shattered by the process above)?

Your Answer		Score	Explanation
$d + 1$			
∞	✔	20.00	
$N + d + 1$			
1			
$N + 1$			
Total		20.00 / 20.00	

Question 13

For Questions 13-15, you will play with linear regression and feature transforms. Consider the target function:

$$f(x_1, x_2) = \text{sign}(x_1^2 + x_2^2 - 0.6)$$

Generate a training set of $N = 1000$ points on $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability of picking each $\mathbf{x} \in \mathcal{X}$. Generate simulated noise by flipping the sign of the output in a random 10% subset of the generated training set.

Carry out Linear Regression without transformation, i.e., with feature vector: $(1, x_1, x_2)$, to find the weight $\mathbf{w}$, and use $\mathbf{w}_{\text{lin}}$ directly for classification. What is the closest value to the classification $(0/1)$ in-sample error (E_{in})? Run the experiment 1000 times and take the average E_{in} in order to reduce variation in your results.

Your Answer		Score	Explanation
0.7			
0.9			
0.5	✔	20.00	
0.1			
0.3			
Total		20.00 / 20.00	

Question 14

Now, transform the training data into the following nonlinear feature vector:

$$(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

Find the vector $\tilde{\mathbf{w}}$ that corresponds to the solution of Linear Regression, and take it for classification.

Which of the following hypotheses is closest to the one you find using Linear Regression on the transformed input? Closest here means agrees the most with your hypothesis (has the most probability of agreeing on a randomly selected point).

Your Answer	Score	Explanation
$g(x_1, x_2) = \text{sign}(-1 - 1.5x_1 + 0.08x_2 + 0.13x_1x_2 + 0.05x_1^2 + 1.5x_2^2)$		
$g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 15x_2^2)$		

◉ $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 1.5x_2^2)$	✔	20.00
○ $g(x_1, x_2) = \text{sign}(-1 - 1.5x_1 + 0.08x_2 + 0.13x_1x_2 + 0.05x_1^2 + 0.05x_2^2)$		
○ $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 15x_1^2 + 1.5x_2^2)$		
Total		20.00 / 20.00

Question 15

What is the closest value to the classification out-of-sample error E_{out} of your hypothesis? Estimate it by generating a new set of 1000 points and adding noise as before. Average over 1000 runs to reduce the variation in your results.

Your Answer		Score	Explanation
○ 0.7			
○ 0.9			
◉ 0.1	✔	20.00	
○ 0.5			
○ 0.3			
Total		20.00 / 20.00	

Question 16

For Questions 16-17, you will derive an algorithm for multinomial (multiclass) logistic regression. For a K-class classification problem, we will denote the output space $\mathcal{Y} = \{1, 2, \cdots, K\}$. The hypotheses considered by MLR are indexed by a list of weight vectors $(\mathbf{w}_1, \cdots, \mathbf{w}_K)$, each weight vector of length $d + 1$. Each list represents a hypothesis

$$h_y(\mathbf{x}) = \frac{\exp(\mathbf{w}_y^T \mathbf{x})}{\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x})}$$

that can be used to approximate the target distribution $P(y|\mathbf{x})$. MLR then seeks for the maximum likelihood solution over all such hypotheses.

For general K, derive an $E_{\text{in}}(\mathbf{w}_1, \cdots, \mathbf{w}_K)$ like page 11 of Lecture 10 slides by minimizing the negative log likelihood.

Your Answer		Score	Explanation
none of the other choices			
$\frac{1}{N} \sum_{n=1}^{N} \left(\ln \left(\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x}_n) - \exp(\mathbf{w}_{y_n}^T \mathbf{x}_n) \right) \right)$			
$\frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{K} \left(\mathbf{w}_i^T \mathbf{x}_n - \mathbf{w}_{y_n}^T \mathbf{x}_n \right) \right)$			
$\frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{K} \mathbf{w}_i^T \mathbf{x}_n - \mathbf{w}_{y_n}^T \mathbf{x}_n \right)$			
$\frac{1}{N} \sum_{n=1}^{N} \left(\ln \left(\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x}_n) \right) - \mathbf{w}_{y_n}^T \mathbf{x}_n \right)$	✔	20.00	
Total		20.00 / 20.00	

Question 17

For the E_{in} derived above, its gradient ∇E_{in} can be represented by $\left(\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_1}, \frac{\partial E_{\text{in}}}{\partial \mathbf{w}_2}, \cdots, \frac{\partial E_{\text{in}}}{\partial \mathbf{w}_K} \right)$, write down $\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_i}$.

Your Answer		Score	Explanation
$\frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{K} \left(\exp(\mathbf{w}_i^T \mathbf{x}_n) - [y_n = i] \right) \mathbf{x}_n \right)$			
$\frac{1}{N} \sum_{n=1}^{N} \left(\left(h_i(\mathbf{x}_n) - [y_n = i] \right) \mathbf{x}_n \right)$	✔	20.00	
none of the other choices			
$\frac{1}{N} \sum_{n=1}^{N} \left(\left(h_i(\mathbf{x}_n) - 1 \right) \mathbf{x}_n \right)$			
$\frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{K} \left(\exp(\mathbf{w}_i^T \mathbf{x}_n) - 1 \right) \mathbf{x}_n \right)$			
Total		20.00 / 20.00	

Question 18

For Questions 18-20, you will play with logistic regression.

Please use the following set for training:

https://d396qusza40orc.cloudfront.net/ntumlone%2Fhw3%2Fhw3_train.dat

and the following set for testing:

https://d396qusza40orc.cloudfront.net/ntumlone%2Fhw3%2Fhw3_test.dat

Implement the fixed learning rate gradient descent algorithm for logistic regression. Run the algorithm with $\eta = 0.001$ and $T = 2000$, what is $E_{out}(g)$ from your algorithm, evaluated using the $0/1$ error on the test set?

Your Answer		Score	Explanation
0.103			
0.475	✔	20.00	
0.220			
0.412			
0.322			
Total		20.00 / 20.00	

Question 19

Implement the fixed learning rate gradient descent algorithm for logistic regression. Run the algorithm with $\eta = 0.01$ and $T = 2000$, what is $E_{out}(g)$ from your algorithm, evaluated using the $0/1$ error on the test set?

Your Answer	Score	Explanation

0.322		
0.103		
0.220	✔	20.00
0.475		
0.412		
Total		20.00 / 20.00

Question 20

Implement the fixed learning rate stochastic gradient descent algorithm for logistic regression. Instead of randomly choosing n in each iteration, please simply pick the example with the cyclic order $n = 1, 2, \ldots, N, 1, 2, \ldots$.

Run the algorithm with $\eta = 0.001$ and $T = 2000$, what is $E_{out}(g)$ from your algorithm, evaluated using the $0/1$ error on the test set?

Your Answer		Score	Explanation
0.412			
0.475	✔	20.00	
0.219			
0.105			
0.328			
Total		20.00 / 20.00	