作業三

TOTAL POINTS 200

---

1. **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:** 10 points

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

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?

- 10
- 25
- **100**
- 500
- 1000

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 $\mathrm{X} \in \mathbb{R}^{N \times (d+1)}$ containing N examples with d features. Assume $\mathrm{X}^T\mathrm{X}$ is invertible and $N > d + 1$, which statement of H is true?** 10 points

- H is always invertible.
- $(d+1)$ eigenvalues of H are bigger than 1.
- none of the other choices
- **$\mathrm{H}^{1126} = \mathrm{H}$.**
- $N - (d+1)$ eigenvalues of H are 1.

3. **Which of the following is an upper bound of $[[\text{sign}(\mathbf{w}^T\mathbf{x}) \neq y]]$ for $y \in \{-1, +1\}$?** 10 points

- none of the other choices

- $err(\mathbf{w}) = [[\mathbf{w}^T\mathbf{x} \geq y]]$
- $err(\mathbf{w}) = \frac{1}{2}\exp(-y\mathbf{w}^T\mathbf{x})$
- (selected) $err(\mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$
- $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$

4. Which of the following is a differentiable function of $\mathbf{w}$ everywhere? 10 points

- $err(\mathbf{w}) = [[\mathbf{w}^T\mathbf{x} \geq y]]$
- (selected) $err(\mathbf{w}) = \frac{1}{2}\exp(-y\mathbf{w}^T\mathbf{x})$
- $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$
- $err(\mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$
- none of the other choices

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? 10 points

- $err(\mathbf{w}) = \frac{1}{2}\exp(-y\mathbf{w}^T\mathbf{x})$
- $err(\mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$
- $err(\mathbf{w}) = [[\mathbf{w}^T\mathbf{x} \geq y]]$
- (selected) $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$
- none of the other choices

6. For Questions 6-10, 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 $(u, v) = (0, 0)$? 10 points

- (selected) $(-2, 0)$
- $(3, -1)$
- none of the other choices
- $(0, -2)$
- $(-3, 1)$

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?** 10 points

- 4.904
- 3.277
- 2.825
- 2.361
- 1.436

8. **Continuing 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** 10 points

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

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

- none of the other choices
- $(3, 8, -0.5, -1, -2, 0)$
- $(1.5, 4, -1, -2, 0, 3)$
- $(3, 8, -1, -2, 0, 3)$
- $(1.5, 4, -0.5, -1, -2, 0)$

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.)** 10 points

- none of the other choices
- $-(\nabla^2 E(u, v))^{-1} \nabla E(u, v)$
- $-\nabla^2 E(u, v) \nabla E(u, v)$
- $+\nabla^2 E(u, v) \nabla E(u, v)$
- $+(\nabla^2 E(u, v))^{-1} \nabla E(u, v)$

10. **Use the Newton direction (without \eta) for updating and start from $(u_0, v_0) = (0, 0)$.** 10 points

What is $E(u_5, v_5)$ after five updates?

- 4.904
- 3.277
- 2.825
- 2.361 (selected)
- 1.436

11. 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 $\mathbf{x}$? 10 points

- $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$
- $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$ (selected)
- $\mathbf{x}_1, \mathbf{x}_3$
- $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$
- $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$

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 10 points

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

Consider a learning algorithm that performs PLA after the feature transform. Assume that all $\mathbf{x}_n$ are different, 30% of the y_n's are positive, and $\text{sign}(0) = +1$. Then, estimate the E_{out} of the algorithm with a test set with all its input vectors $\mathbf{x}$ different from those training $\mathbf{x}_n$'s and 30% of its output labels y to be positive. Which of the following is not true?

- $E_{out} = 0.7$
- PLA will halt after enough iterations.
- $E_{in} = 0.7$ (selected)
- The transformed data set is always linearly separable in the $\mathcal{Z}$ space.
- All $\mathbf{z}_n$'s are orthogonal to each other.

13. **For Questions 13-15, consider the target function:** 10 points

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

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 $\mathbf{w}_{\text{lin}}$, 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.

- ○ 0.1
- ○ 0.3
- ◉ 0.5
- ○ 0.7
- ○ 0.9

14. **Now, transform the training data into the following nonlinear feature vector:** 10 points

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

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).

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

15. **Following Question 14, 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** 10 points

results.

- (●) 0.1
- ( ) 0.3
- ( ) 0.5
- ( ) 0.7
- ( ) 0.9

16. **For Questions 16-17, you will derive an algorithm for the multinomial (multiclass) logistic regression model. For a K-class classification problem, we will denote the output space $\mathcal{Y} = \{1, 2, \dots, K\}$. The hypotheses considered by the model are indexed by a list of weight vectors $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K)$, each weight vector of length $d+1$. Each list represents a hypothesis** 10 points

$$h_y(\mathbf{x}) = \Big(\exp(\mathbf{w}_y^T \mathbf{x})\Big) / \Big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x})\Big)$$

that can be used to approximate the target distribution $P(y|\mathbf{x})$. The model 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. What is the resulting E_{in}?

- ( ) $\frac{1}{N} \sum_{n=1}^{N} \Big(\ln\Big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x}_n) - \exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)\Big)\Big)$
- (●) $\frac{1}{N} \sum_{n=1}^{N} \Big(\ln\Big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x}_n)\Big) - \mathbf{w}_{y_n}^T \mathbf{x}_n\Big)$
- ( ) none of the other choices
- ( ) $\frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{k=1}^{K} \mathbf{w}_k^T \mathbf{x}_n - \mathbf{w}_{y_n}^T \mathbf{x}_n\Big)$
- ( ) $\frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{k=1}^{K} \big(\mathbf{w}_k^T \mathbf{x}_n - \mathbf{w}_{y_n}^T \mathbf{x}_n\big)\Big)$

17. **For the E_{in} derived above, its gradient ∇E_{in} can be represented by $\Big(\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_1}, \frac{\partial E_{\text{in}}}{\partial \mathbf{w}_2}, \cdots, \frac{\partial E_{\text{in}}}{\partial \mathbf{w}_K}\Big)$, write down $\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_i}$.** 10 points

- ( ) none of the other choices
- ( ) $\frac{1}{N} \sum_{n=1}^{N} \big(\big(h_i(\mathbf{x}_n) - [[y_n = i]]\big)\mathbf{x}_n\big)$
- ( ) $\frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{i=1}^{K} \big(\exp(\mathbf{w}_i^T \mathbf{x}_n) - [[y_n = i]]\big)\mathbf{x}_n\Big)$

- $\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{i=1}^{K}\left(\exp(\mathbf{w}_i^T\mathbf{x}_n)-1\right)\mathbf{x}_n\right)$
- $\frac{1}{N}\sum_{n=1}^{N}\left(\left(h_i(\mathbf{x}_n)-1\right)\mathbf{x}_n\right)$

18. **For Questions 18-20, you will play with logistic regression. Please use the following set for training:** 10 points

https://www.csie.ntu.edu.tw/~htlin/mooc/datasets/mlfound_algo/hw3_train.dat

and the following set for testing:

https://www.csie.ntu.edu.tw/~htlin/mooc/datasets/mlfound_algo/hw3_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?

- 0.475
- 0.412
- 0.322
- 0.220
- 0.103

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?** 10 points

- 0.475
- 0.412
- 0.322
- 0.220
- 0.103

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 $n = 1, 2, \ldots, N, 1, 2, \ldots$.** 10 points

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?

- 0.475
- 0.412
- 0.322
- 0.220
- 0.103