Homework Assignment 8

Due: Friday, December 5, 2025, 11:59 p.m. Mountain time
Total marks: 15

Policies:

For all multiple-choice questions, note that multiple correct answers may exist. However, selecting an incorrect option will cancel out a correct one. For example, if you select two answers, one correct and one incorrect, you will receive zero points for that question. Similarly, if the number of incorrect answers selected exceeds the correct ones, your score for that question will be zero. Please note that it is not possible to receive negative marks. **You must select all the correct options to get full marks for the question.**
This PDF version of the questions has been provided for your convenience should you wish to print them and work offline.
Only answers submitted through the Canvas quiz system will be graded. Please do not submit a written copy of your responses.

Question 1. [1 MARK]

Let $\hat{f}_{\text{ERM}}$ be as defined in section 9.1.1 of the course notes. Which of the following is true?

a. We can think of $\hat{f}_{\text{ERM}}(\mathbf{x})$ as predicting the probability of $\mathbf{x}$ belonging to class 1.

b. It is always the case that f_{Bayes} outputs class 1 if $\hat{f}_{\text{ERM}}(\mathbf{x}) \geq 0.5$, and class 0 otherwise.

c. f_{Bayes} is equal to $\hat{f}_{\text{ERM}}$.

d. $\hat{f}_{\text{ERM}}$ has the same closed-form solution as the ERM predictor for linear regression with the squared loss.

Question 2. [1 MARK]

In class we used the following function class for logistic regression:

$$\mathcal{F} = \left\{ f \mid f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f(\mathbf{x}) = \sigma(\mathbf{x}^\top \mathbf{w}), \text{ and } \mathbf{w} \in \mathbb{R}^{d+1} \right\}.$$

Suppose that we would like to use a larger function class that contains polynomial features of the input $\mathbf{x}$. Recall that $\phi_p(\mathbf{x})$ is the degree p polynomial feature map of $\mathbf{x}$. We define the new function class as follows

$$\mathcal{F}_p = \left\{ f \mid f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f(\mathbf{x}) = \sigma(\phi_p(\mathbf{x})^\top \mathbf{w}), \text{ and } \mathbf{w} \in \mathbb{R}^{\bar{p}} \right\},$$

where $\bar{p} = \binom{d+p}{p}$ is the number of features in the polynomial feature map $\phi_p(\mathbf{x})$. Is the following statement true or false? For all $p \in \{2, \dots\}$ it holds that $\mathcal{F} \subset \mathcal{F}_p \subset \mathcal{F}_{p+1}$.

Question 3. [1 MARK]

Let everything be as defined in the previous question. Let $\hat{f}_{\text{ERM},p}(\mathbf{x}) = \sigma(\phi_p(\mathbf{x})^\top \mathbf{w}_{\text{ERM},p})$ be the ERM predictor for the function class $\mathcal{F}_p$, where $\mathbf{w}_{\text{ERM},p}$ is the minimizer of the estimated loss (with the binary cross-entropy loss function). Is the following statement true or false? The binary predictor $\hat{f}_{\text{Bin}}$ that outputs class 1 if $\hat{f}_{\text{ERM},p}(\mathbf{x}) \geq 0.5$ and class 0 otherwise can be equivalently defined as

$$\hat{f}_{\text{Bin}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \phi_p(\mathbf{x})^\top \mathbf{w}_{\text{ERM},p} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Question 4. [1 MARK]

In class we worked out the MLE solution for binary classification. In this question we are going to work out the MAP solution. Suppose that the setting is the same as in the MLE setting defined in section 9.1 of the course notes. However, we will also assume that the weights $w_1^*, \ldots, w_d^*$ are i.i.d. Gaussian random variables with mean 0 and variance $1/\lambda$. The bias term w_0^* is also independent of the other weights and has a uniform distribution on $[-a, a]$, for a very large a. Which of the following is equal to $\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w} \in \mathbb{R}^{d+1}} p(\mathbf{w} \mid \mathcal{D})$?

a.

$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left[-\sum_{i=1}^{n} \left(y_i \log\left(\sigma(\mathbf{x}_i^\top \mathbf{w})\right) + (1 - y_i) \log\left(1 - \sigma(\mathbf{x}_i^\top \mathbf{w})\right) \right) + \frac{\lambda}{2} \sum_{j=1}^{d} w_j^2 \right]$$

b.

$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left[-\frac{\lambda}{2} \sum_{i=1}^{n} \left(\left(y_i \log\left(\sigma(\mathbf{x}_i^\top \mathbf{w})\right) + (1 - y_i) \log\left(1 - \sigma(\mathbf{x}_i^\top \mathbf{w})\right) \right) \left(\sum_{j=1}^{d} w_j^2 \right) \right) \right]$$

c.

$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left[\sum_{i=1}^{n} \left(y_i - \sigma(\mathbf{x}_i^\top \mathbf{w}) \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{d} w_j^2 \right]$$

d.

$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left[-\sum_{i=1}^{n} \left(y_i \log\left(\sigma(\mathbf{x}_i^\top \mathbf{w})\right) + (1 - y_i) \log\left(1 - \sigma(\mathbf{x}_i^\top \mathbf{w})\right) \right) + \lambda \sum_{j=1}^{d} w_j \log(w_j) \right]$$

Question 5. [1 MARK]

In this question, we are going to work out the MAP solution for binary classification using a Laplace prior. Suppose that the setting is the same as in the MLE setting defined in section 9.1 of the course notes. However, we will also assume that the weights $w_1^*, \ldots, w_d^*$ are i.i.d. Laplace random variables with mean 0 and scale parameter $1/\lambda$. The bias term w_0^* is also independent of the other weights and has a uniform distribution on $[-a, a]$ for a very large a.
Which of the following is equal to $\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w} \in \mathbb{R}^{d+1}} p(\mathbf{w} \mid \mathcal{D})$?

a.

$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left[-\sum_{i=1}^{n} \left(y_i \log\left(\sigma(\mathbf{x}_i^\top \mathbf{w})\right) + (1 - y_i) \log\left(1 - \sigma(\mathbf{x}_i^\top \mathbf{w})\right) \right) + \lambda \sum_{j=1}^{d} |w_j| \right]$$

b.
$$\arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left[-\sum_{i=1}^{n} \Big(y_i \log\Big(\sigma(\mathbf{x}_i^\top \mathbf{w})\Big) + (1-y_i)\log\Big(1-\sigma(\mathbf{x}_i^\top \mathbf{w})\Big)\Big) + \lambda \sum_{j=0}^{d} |w_j| \right]$$

c.
$$\arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left[\sum_{i=1}^{n} \Big(y_i - \sigma(\mathbf{x}_i^\top \mathbf{w})\Big)^2 + \lambda \sum_{j=1}^{d} |w_j| \right]$$

d.
$$\arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left[-\sum_{i=1}^{n} \Big(y_i \log\Big(\sigma(\mathbf{x}_i^\top \mathbf{w})\Big) + (1-y_i)\log\Big(1-\sigma(\mathbf{x}_i^\top \mathbf{w})\Big)\Big) + \lambda \sum_{j=1}^{d} \log(|w_j|) \right]$$

Question 6. [1 MARK]

Suppose that things are as defined in the previous question. However, we assume the weights $w_0^*, w_1^*, \ldots, w_d^*$ are all i.i.d. uniform random variables on $[-a, a]$ for a very large a. Is the following statement true or false? The MAP solution with this prior is equivalent to the MLE solution.

Question 7. [1 MARK]

Let $\hat{f}_{\text{Mul}}$ be as defined in section 9.2 of the course notes. Which of the following is true?

a. $\hat{f}_{\text{Mul}}$ outputs a vector of probabilities where the y-th element is the probability of class y.

b. $\sigma(\mathbf{x}^\top \mathbf{w}_{\text{MLE},0}, \ldots, \mathbf{x}^\top \mathbf{w}_{\text{MLE},K-1})$ outputs a vector of probabilities where the y-th element is the probability of class y.

c. $\hat{f}_{\text{ERM}}$ as defined in section 9.2.1 of the course notes is approximately equal to f_{Bayes}.

d. There is no closed-form solution for $\mathbf{w}_{\text{MLE},k}$ for any k.

Question 8. [1 MARK]

Let everything be as defined in the previous question. Which of the following is true?

a. If $\mathbf{x}^\top \mathbf{w}_{\text{MLE},y} < \mathbf{x}^\top \mathbf{w}_{\text{MLE},k}$ for all $k \neq y$, then $\hat{f}_{\text{Mul}}(\mathbf{x}) = y$.

b. If $\mathbf{x}^\top \mathbf{w}_{\text{MLE},y} > \mathbf{x}^\top \mathbf{w}_{\text{MLE},k}$ for all $k \neq y$, then $\hat{f}_{\text{Mul}}(\mathbf{x}) = y$.

c. If $\sigma_y(\mathbf{x}^\top \mathbf{w}_{\text{MLE},0}, \ldots, \mathbf{x}^\top \mathbf{w}_{\text{MLE},K-1}) > 0.5$, then $\hat{f}_{\text{Mul}}(\mathbf{x}) = y$.

d. If $\mathbf{x}^\top(\mathbf{w}_{\text{MLE},y} - \mathbf{w}_{\text{MLE},k}) = 0$, then
$$p(y \mid \mathbf{x}, \mathbf{w}_{\text{MLE},0}, \ldots, \mathbf{w}_{\text{MLE},K-1}) = p(k \mid \mathbf{x}, \mathbf{w}_{\text{MLE},0}, \ldots, \mathbf{w}_{\text{MLE},K-1}).$$

Question 9. [1 MARK]

Let $\mathcal{Y} = \{0, 1\}$ be the set of labels. Define the following two label functions:

$$h(y) = \begin{cases} (1,0)^\top & \text{if } y = 0, \\ (0,1)^\top & \text{if } y = 1. \end{cases}$$

$$r(\hat{y}) = (1 - \hat{y}, \hat{y})^\top.$$

Is the following statement true or false? For any $\hat{y} \in (0, 1)$ and $y \in \mathcal{Y}$ (where $(0, 1)$ is the open interval from 0 to 1), the binary cross-entropy loss with input $\hat{y}$ and y is equal to the multiclass cross-entropy loss with input $r(\hat{y})$ and $h(y)$.

Question 10. [1 MARK]

Suppose you are in the binary classification setting as defined in section 9.1 of the course notes and you solve for $\mathbf{w}_{\text{MLE}}$. Now suppose that the setting is the multiclass classification setting (with $K = 2$) as defined in section 9.2 of the course notes and you solve for $\mathbf{w}_{\text{MLE},0}$, $\mathbf{w}_{\text{MLE},1}$. Is the following true or false? The solution for $\mathbf{w}_{\text{MLE}}$ in the binary classification setting is the same as the solution for $\mathbf{w}_{\text{MLE},1}$ in the multiclass classification setting.

Question 11. [1 MARK]

You are designing a neural network architecture for a binary classification problem. You decide to have $B = 5$ layers and $d^{(1)} = 50$, $d^{(2)} = 40$, $d^{(3)} = 30$, $d^{(4)} = 20$, $d^{(5)} = 1$ neurons in each layer respectively. The input dimension is $d = 100$. How many weight vectors are there in the network?

Question 12. [1 MARK]

Let everything be as defined in the previous question. If you sum up the dimension of all the weight vectors in the neural network you get the number of weights in the network. How many weights are there in the network?

Question 13. [1 MARK]

The tanh function is defined as $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$, where $z \in \mathbb{R}$. Is the following statement true or false? The tanh function is a valid activation function?

Question 14. [1 MARK]

The logistic function is defined as $\sigma(z) = \frac{1}{1+e^{-z}}$, where $z \in \mathbb{R}$. You would like to define the function $f(\mathbf{x}) = \sigma(\mathbf{x}^\top \mathbf{w})$ where $\mathbf{x} \in \mathbb{R}^{d+1}, \mathbf{w} \in \mathbb{R}^{d+1}$ as a neural network. Which of the following is correct?

a. The neural network has $B = 2$ layers, $d^{(1)} = 1, d^{(2)} = 1$ neurons, activation $h = \sigma$, and two weight vectors $\mathbf{w}_1^{(1)} = \mathbf{w}, \mathbf{w}_1^{(2)} = (1, 1)^\top$.

b. The neural network has $B = 1$ layer, $d^{(1)} = 1$ neuron, activation $h = \sigma$, and one weight vector $\mathbf{w}_1^{(1)} = \mathbf{w}$.

c. The neural network has $B = 1$ layer, $d^{(1)} = 1$ neuron, activation $h(z) = z$, and one weight vector $\mathbf{w}_1^{(1)} = \mathbf{w}$.

d. The neural network has $B = 2$ layers, $d^{(1)} = 1, d^{(2)} = 1$ neurons, activation $h = \sigma$ in the first layer, activation $h(z) = z$ in the second layer, and two weight vectors $\mathbf{w}_1^{(1)} = \mathbf{w}, \mathbf{w}_1^{(2)} = (0, 1)^\top$.

Question 15. [1 MARK]

You have a neural network f with $B = 2$ layers and $d^{(1)} = 3$, $d^{(2)} = 1$ neurons in each layer respectively. The input dimension is $d = 2$. You choose to use the ReLU activation function, defined as $\text{ReLU}(z) = \max(0, z)$, where $z \in \mathbb{R}$. The weight vectors have the following values:

$$\mathbf{w}_1^{(1)} = (1, 1, 1)^\top \quad \mathbf{w}_2^{(1)} = (-1, -1, -1)^\top \quad \mathbf{w}_3^{(1)} = (-1, 0, 1)^\top \quad \mathbf{w}_1^{(2)} = (1, 1, 1, 1)^\top$$

Suppose you get a feature vector $\mathbf{x} = (1, -1, 1)^\top$. What is $f(\mathbf{x})$?