Homework Assignment 6

Due: Friday, November 8, 2024, 11:59 p.m. Mountain time

Total marks: 26

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.**

While the syllabus initially indicated the need to submit a paragraph explaining the use of AI or other resources in your assignments, this requirement no longer applies as we are now utilizing eClass quizzes instead of handwritten submissions. Therefore, you are **not** required to submit any explanation regarding the tools or resources (such as online tools or AI) used in completing this quiz.

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 eClass quiz system will be graded. Please do not submit a written copy of your responses.

Question 1. [1 MARK]

Consider the predictor $f(x) = xw$, where $w \in \mathbb{R}$ is a one-dimensional parameter, and x represents the feature with no bias term. Suppose you are given a dataset of n data points $\mathcal{D} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, where each y_i is the target variable corresponding to feature x_i. Let the loss function be the scaled squared loss $\ell(f(x), y) = c(f(x) - y)^2$ where $c \in \mathbb{R}$. The estimate of the expected loss for a parameter $w \in \mathbb{R}$ is defined as the following convex function:

$$\hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} c(x_i w - y_i)^2$$

What is the closed form solution for $\hat{w} = \arg\min_{w \in \mathbb{R}} \hat{L}(w)$?

a. $\hat{w} = \frac{\sum_{i=1}^{n} c x_i y_i}{\sum_{i=1}^{n} x_i^2}$

b. $\hat{w} = \frac{\sum_{i=1}^{n} y_i}{n}$

c. $\hat{w} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$

d. $\hat{w} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Solution:

Answer: d.

Explanation: To find the closed-form solution for $\hat{w}$, we need to minimize $\hat{L}(w)$. This is equivalent to minimizing the function:

$$\hat{L}(w) = \frac{c}{n}\sum_{i=1}^{n}(x_i w - y_i)^2$$

Taking the derivative with respect to w and setting it to zero gives:

$$\frac{\partial \hat{L}}{\partial w} = \frac{c}{n}\sum_{i=1}^{n} 2(x_i w - y_i)x_i = 0$$

Simplifying leads to:

$$\sum_{i=1}^{n} x_i(x_i w) = \sum_{i=1}^{n} x_i y_i$$

Thus, solving for w yields:

$$\hat{w} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Question 2. [1 MARK]

Let everything be defined as in the previous question. Suppose we consider the multivariate case where $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$, and $\mathbf{w} \in \mathbb{R}^{d+1}$. What is the closed form solution for $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \hat{L}(\mathbf{w})$?

a. $\hat{\mathbf{w}} = A^{-1}b$ where $A = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ and $b = \sum_{i=1}^{n} \mathbf{x}_i y_i$ (assume that A is invertible).

b. $\hat{\mathbf{w}} = Ax$ where $A = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$

c. $\hat{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$

d. $\hat{\mathbf{w}} = \frac{\sum_{i=1}^{n} c\mathbf{x}_i y_i}{\sum_{i=1}^{n} c\mathbf{x}_i^2}$

Solution:

Answer: a.
Explanation: To find the closed-form solution for $\hat{\mathbf{w}}$, we minimize the expected loss defined as:

$$\hat{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} c(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

Taking the derivative with respect to $\mathbf{w}$ and setting it to zero gives:

$$\frac{\partial \hat{L}}{\partial \mathbf{w}} = \frac{c}{n}\sum_{i=1}^{n} 2(\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i = 0$$

This simplifies to:

$$\sum_{i=1}^{n} \mathbf{x}_i(\mathbf{x}_i^\top \mathbf{w}) = \sum_{i=1}^{n} y_i \mathbf{x}_i$$

We can express this using the definitions A and b:

$$A\mathbf{w} = b$$

Thus, we can write:

$$\hat{\mathbf{w}} = A^{-1}b$$

Question 3. [1 MARK]

Let $g(w) = -\ln w \sum_{i=1}^n y_i - \ln(1-w)\sum_{i=1}^n (1-y_i)$ where $w \in \mathbb{R}$. We can rewrite this a bit more simply as $g(w) = -s\ln w - (n-s)\ln(1-w)$ where $s = \sum_{i=1}^n y_i$. What is the derivative $g'(w)$ and the first order gradient descent update rule with a constant step size η?

a. $g'(w) = -\frac{s}{1-w} + \frac{n-s}{w}$ and update rule $w \leftarrow w - \eta\left(-\frac{s}{1-w} + \frac{n-s}{w}\right)$

b. $g'(w) = -\frac{s}{w} + \frac{n-s}{1-w}$ and update rule $w \leftarrow w - \eta\left(-\frac{s}{1-w} + \frac{n-s}{w}\right)$

c. $g'(w) = -\frac{s}{w} + \frac{n-s}{1-w}$ and update rule $w \leftarrow w - \eta\left(-\frac{s}{w} + \frac{n-s}{1-w}\right)$

d. $g'(w) = -\frac{s}{1-w} - \frac{n-s}{w}$ and update rule $w \leftarrow w - \eta\left(-\frac{s}{1-w} - \frac{n-s}{w}\right)$

Solution:

Answer: c
Explanation: To find the derivative $g'(w)$, we differentiate $g(w)$:

$$g(w) = -s\ln w - (n-s)\ln(1-w)$$

Taking the derivative:

$$g'(w) = -s\frac{1}{w} + (n-s)\frac{1}{1-w}$$

The gradient descent update rule with a constant step size η is given by:

$$w \leftarrow w - \eta g'(w) = w - \eta\left(-\frac{s}{w} + \frac{n-s}{1-w}\right)$$

Thus, the first-order gradient descent update rule becomes:

$$w \leftarrow w + \eta\left(\frac{s}{w} - \frac{n-s}{1-w}\right)$$

Question 4. [1 MARK]

Let everything be defined as in the previous question. What is the second derivative $g''(w)$ and the second order gradient descent update rule?

a. $g''(w) = \frac{s}{w^2} - \frac{n-s}{(1-w)^2}$ and update: $w \leftarrow w - \frac{-\frac{s}{w}+\frac{n-s}{1-w}}{\frac{s}{w^2}-\frac{n-s}{(1-w)^2}}$

b. $g''(w) = \frac{s}{w^2} + \frac{n-s}{(1-w)^2}$ and update: $w \leftarrow w - \frac{-\frac{s}{w}+\frac{n-s}{1-w}}{\frac{s}{w^2}+\frac{n-s}{(1-w)^2}}$

c. $g''(w) = -\frac{s}{w^2} + \frac{n-s}{(1-w)^2}$ and update: $w \leftarrow w - \frac{-\frac{s}{w}+\frac{n-s}{1-w}}{-\frac{s}{w^2}+\frac{n-s}{(1-w)^2}}$

d. $g''(w) = \frac{s}{w^2} + \frac{n-s}{(1-w)^2}$ and update: $w \leftarrow w + \frac{-\frac{s}{w}+\frac{n-s}{1-w}}{\frac{s}{w^2}+\frac{n-s}{(1-w)^2}}$

Solution:

Answer: b.
Explanation: To find the second derivative:

$$g''(w) = \frac{s}{w^2} + \frac{n-s}{(1-w)^2}$$

The first derivative is given by:

$$g'(w) = -\frac{s}{w} + \frac{n-s}{1-w}$$

Thus, the second-order gradient descent update rule is:

$$w \leftarrow w - \frac{g'(w)}{g''(w)} = w - \frac{-\frac{s}{w} + \frac{n-s}{1-w}}{\frac{s}{w^2} + \frac{n-s}{(1-w)^2}}$$

Question 5. [1 MARK]

Let everything be defined as in the previous question. What is the closed form solution for

$$w^* = \arg\min_{w \in \mathbb{R}} g(w)$$

a. $w^* = n/s$

b. $w^* = s/(n-s)$

c. $w^* = s/(s-n)$

d. $w^* = s/n$

Solution:

Answer: d.
Explanation: Set derivative of $g(w)$ to zero and solve for w:

$$\frac{-s}{w} + \frac{n-s}{1-w} = 0 \implies \frac{s}{w} = \frac{n-s}{1-w} \implies w = \frac{s}{n}\,.$$

Question 6. [1 MARK]

Let $g(w) = w^4 + e^{-w}$ where $w \in \mathbb{R}$. What is the derivative $g'(w)$ and the first order gradient descent update rule with a constant step size η?

a. $g'(w) = 4w^3 - e^{-w}$ and update: $w \leftarrow w - \eta(4w^3 - e^{-w})$

b. $g'(w) = 4w^3 + e^{-w}$ and update: $w \leftarrow w - \eta(4w^3 + e^{-w})$

c. $g'(w) = 4w^3 + e^{-w}$ and update: $w \leftarrow w + \eta(4w^3 + e^{-w})$

d. $g'(w) = 4w^3 - e^{-w}$ and update: $w \leftarrow w + \eta(4w^3 - e^{-w})$

Solution:

Answer: a.
Explanation: To find the derivative $g'(w)$:

$$g'(w) = \frac{d}{dw}(w^4) + \frac{d}{dw}(e^{-w}) = 4w^3 - e^{-w}.$$

The first-order gradient descent update rule is given by:

$$w \leftarrow w - \eta g'(w) = w - \eta(4w^3 - e^{-w}).$$

Question 7. [1 MARK]

Let everything be defined as in the previous question. What is the second derivative $g''(w)$ and the second order gradient descent update rule?

a. $g''(w) = 12w^2 - e^{-w}$ and update: $w \leftarrow w - \frac{4w^3 - e^{-w}}{12w^2 - e^{-w}}$

b. $g''(w) = 12w^2 + e^{-w}$ and update: $w \leftarrow w + \frac{4w^3 - e^{-w}}{12w^2 + e^{-w}}$

c. $g''(w) = 12w^2 + e^{-w}$ and update: $w \leftarrow w - \frac{4w^3 - e^{-w}}{12w^2 + e^{-w}}$

d. $g''(w) = 12w^2 - e^{-w}$ and update: $w \leftarrow w + \frac{4w^3 - e^{-w}}{12w^2 - e^{-w}}$

Solution:

Answer: c.
Explanation: To find the second derivative $g''(w)$:
1. First, we have the first derivative:

$$g'(w) = 4w^3 - e^{-w}.$$

2. Next, we differentiate $g'(w)$ to get $g''(w)$:

$$g''(w) = \frac{d}{dw}(4w^3) - \frac{d}{dw}(e^{-w}) = 12w^2 + e^{-w}.$$

The second-order gradient descent update rule is given by:

$$w \leftarrow w - \frac{g'(w)}{g''(w)} = w - \frac{4w^3 - e^{-w}}{12w^2 + e^{-w}}.$$

Question 8. [1 MARK]

Let everything be defined as in the previous question. For the second order update rule, calculate $w^{(1)}$ if $w^{(0)} = 0$.

Solution:

Answer: 1.
Explanation: Let $g(w) = w^4 + e^{-w}$ where $w \in \mathbb{R}$. We want to compute $w^{(1)}$ using the second-order gradient descent update rule, given $w^{(0)} = 0$.
The second-order update rule is given by:

$$w^{(1)} = w^{(0)} - \frac{g'(w^{(0)})}{g''(w^{(0)})}$$

Step 1: Calculate $g'(w)$. The first derivative is:

$$g'(w) = 4w^3 - e^{-w}$$

Substituting $w^{(0)} = 0$:

$$g'(0) = 4(0)^3 - e^0 = -1$$

Step 2: Calculate $g''(w)$. The second derivative is:

$$g''(w) = 12w^2 + e^{-w}$$

Substituting $w^{(0)} = 0$:

$$g''(0) = 12(0)^2 + e^0 = 1$$

Step 3: Update the value of w. Now we can compute $w^{(1)}$:

$$w^{(1)} = 0 - \frac{-1}{1} = 0 + 1 = 1$$

Question 9. [1 MARK]

Let everything be defined as in the previous question. Change the step size to be calculated using the normalized gradient. For the first order update rule, calculate $w^{(1)}$ if $w^{(0)} = 0$, $\eta = 1$. Only for this problem, set $\epsilon = 0$.

Solution:

Answer: 1.
Explanation: We know that the derivative at $w^{(0)} = 0$ is $g'(0) = -1$. The normalized gradient step size η is given by

$$\eta^{(0)} = \frac{\eta}{|g'(w^{(0)})|} = 1$$

Therefore

$$w^{(1)} = w^{(0)} - \eta^{(0)} g'(w^{(0)}) = 0 - 1 \times (-1) = 1.$$

Question 10. [1 MARK]

Let $g(\mathbf{w}) = g(w_1, w_2) = w_1^2 w_2^2 + e^{-w_1} + e^{-w_2}$ where $\mathbf{w} \in \mathbb{R}^2$. What is the gradient of $g(w)$ and the first order gradient descent update rule with a constant step size η?

a. $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \left(2w_1^{(t)}(w_2^{(t)})^2, 2w_2^{(t)}(w_1^{(t)})^2\right)^\top$

b. $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \left(2w_1^{(t)}(w_2^{(t)})^2 - e^{-w_1^{(t)}}, 2w_2^{(t)}(w_1^{(t)})^2 - e^{-w_2^x(t)}\right)^\top$

c. $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \left(2w_1^{(t)}(w_2^{(t)})^2 + e^{-w_1^{(t)}}, 2w_2^{(t)}(w_1^{(t)})^2 + e^{-w_2^{(t)}}\right)^\top$

d. $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \left(-2w_1^{(t)}(w_2^{(t)})^2 + e^{-w_1^{(t)}}, -2w_2^{(t)}(w_1^{(t)})^2 + e^{-w_2^{(t)}}\right)^\top$

Solution:

Answer: b
Explanation: Let $g(\mathbf{w}) = g(w_1, w_2) = w_1^2 w_2^2 + e^{-w_1} + e^{-w_2}$ where $\mathbf{w} \in \mathbb{R}^2$.
Step 1: Calculate the gradient $\nabla g(\mathbf{w})$
The gradient is given by:

$$\nabla g(\mathbf{w}) = \left(\frac{\partial g}{\partial w_1}, \frac{\partial g}{\partial w_2}\right)^\top$$

Calculating the partial derivatives:

$$\frac{\partial g}{\partial w_1} = 2w_1 w_2^2 - e^{-w_1}$$

$$\frac{\partial g}{\partial w_2} = 2w_2 w_1^2 - e^{-w_2}$$

Step 2: Write the gradient Thus, the gradient is:

$$\nabla g(\mathbf{w}) = \left(2w_1 w_2^2 - e^{-w_1}, 2w_2 w_1^2 - e^{-w_2}\right)^\top$$

Step 3: Gradient descent update rule The first-order gradient descent update rule is given by:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla g(\mathbf{w}^{(t)})$$

Plugging in the gradient:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \left(2w_1^{(t)}(w_2^{(t)})^2 - e^{-w_1^{(t)}}, 2w_2^{(t)}(w_1^{(t)})^2 - e^{-w_2^{(t)}}\right)^\top$$

Question 11. [1 MARK]

If $\mathcal{F} \subset \mathcal{G}$, then is it true that $\min_{f \in \mathcal{F}} \hat{L}(f) \geq \min_{g \in \mathcal{G}} \hat{L}(g)$?

Solution:

Answer: True.
Explaination: For any $f \in \mathcal{G}$, we know that $\hat{L}(f) \geq \hat{L}(g)$ for all $g \in \mathcal{G}$, since the RHS is the minimum value. But since $\mathcal{F} \subset \mathcal{G}$, we have that $\hat{L}(f) \geq \hat{L}(g)$ for all $f \in \mathcal{F}$ as well. Taking minimum over $\mathcal{F}$ both sides, we have the result.

Question 12. [1 MARK]

Consider the setting of polynomial regression. Let $d = 2$, such that $\mathbf{x} = (x_0 = 1, x_1, x_2)$, and $p = 4$, then $\bar{p} = 10$. True or False?

Solution:

False. It's $\binom{2+4}{4} = 6 \cdot 5/2 = 15$.

Question 13. [1 MARK]

Let everything be defined as in the previous question. The expression for $\phi_p(\mathbf{x})$ is given by

$$\phi(\mathbf{x}) = \left(x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, x_1^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3, x_2^4\right) .$$

True or False?

Solution:

False. The constant term 1 is missing.

Question 14. [1 MARK]

Suppose that

$$\bar{\mathcal{F}}_p = \{f | f : \mathbb{R}^{d+1} \to \mathbb{R},\ \text{and } f(\boldsymbol{x}) = \log(\phi_p(\boldsymbol{x})^\top \boldsymbol{w}),\ \text{for some } \boldsymbol{w} \in \mathbb{R}^{\bar{p}}\}.$$

Is it true that $\bar{\mathcal{F}}_1 \subset \bar{\mathcal{F}}_2$?

Solution:

True. As we increase the degree the function class becomes more expressive.

Question 15. [1 MARK]

You are predicting house prices. Supose you want to make the irriducible error smaller. If you gather a new feature about houses (that you didn't already have) such as the number of swimming pools in the backyard, is it likely to decrease the irriducible error? True or False?

Solution:

True. Irreducible error can be reduced by adding more features that are relevant to the prediction task.

Question 16. [1 MARK]

Consider the same setting as the previous problem. The estimation error can be reduced by reducing the number of data points. True or False?

Solution:

False. Estimation error can be reduced by adding more data points or by using a simpler model.

Question 17. [1 MARK]

Consider the same setting as the previous problem. The approximation error can be reduced by using a larger function class. True or False?

Solution:

True. Approximation error can be reduced by using a more complex model.

Question 18. [1 MARK]

You notice your predictor is overfitting. To reduce overfitting, we should make the degree p of the polynomial function class larger. True or False?

Solution:

False. We need to make p smaller.

Question 19. [1 MARK]

Suppose that you have a small dataset, but a large function class. Would the variance be large or small? Would you expect the bias to be large or small? Would you expect the predictor $\hat{f}_D$ to be underfitting or overfitting the data or neither?

a. variance large, bias large, overfit.

b. variance small, bias large, overfit.

c. variance large, bias small, overfit.

d. variance small, bias small, underfit.

Solution:

Answer: c. Variance large, bias small, overfit.

Question 20. [1 MARK]

Suppose that you have a large dataset, but a small function class, and f_{Bayes} is much more complex than any function in the function class. Would the variance be large or small? Would you expect the bias to be large or small? Would expect the predictor $\hat{f}_D$ to be underfitting or overfitting the data or neither?

a. variance large, bias large, underfit.

b. variance small, bias large, underfit.

c. variance small, bias small, neither overfitting nor underfitting.

d. variance large, bias large, neither overfitting nor underfitting.

Solution:

Answer: b. Variance small, bias large, underfit.

Question 21. [1 MARK]

Suppose that you have a large dataset, a small function class $\mathcal{F}$, and $f_{\text{Bayes}} \in \mathcal{F}$. Would the variance be large or small? Would you expect the bias to be large or small? Would expect the predictor $\hat{f}_D$ to be underfitting or overfitting the data or neither?

a. variance large, bias large, overfitting.

b. variance small, bias small, overfitting.

c. variance small, bias small, neither overfitting nor underfitting.

d. variance large, bias large, neither overfitting nor underfitting.

Solution:

Answer: c. Variance small, bias small, neither overfit or underfit.

Question 22. [1 MARK]

You are using regularization. You notice you are underfitting. You should decrease the value of lambda to reduce underfitting and get a smaller test loss. True or False?

Solution:

True. Decreasing the value of lambda will reduce the regularization strength and allow the model to fit the data better.

Question 23. [1 MARK]

Suppose you have a dataset $\mathcal{D} = (z_1, \ldots, z_n)$ containing n i.i.d. flips of a coin. Since the flips are i.i.d. you know they all follow the distribution Bernoulli (α^*). However, you do not know what α^* is so you would like to estimate it using MLE. Which of the following is the maximum likelihood estimate α_{MLE}?

a. $\alpha_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$

b. $\alpha_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} z_i$

c. $\alpha_{\text{MLE}} = \frac{1}{n-1}\sum_{i=1}^{n} z_i$

d. $\alpha_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n-1} z_i$

Solution:

Answer: b.
Explanation: The probability of each flip z_i is $p(z_i|\alpha) = \alpha^{z_i}(1-\alpha)^{1-z_i}$. The likelihood is:

$$p(\mathcal{D}|\alpha) = \prod_{i=1}^{n} \alpha^{z_i}(1-\alpha)^{1-z_i}$$

The negaitve log-likelihood is:

$$\begin{aligned}
-\log p(\mathcal{D}|\alpha) &= -\sum_{i=1}^{n} \log\left(\alpha^{z_i}(1-\alpha)^{1-z_i}\right) \\
&= -\sum_{i=1}^{n} \left(z_i \log\alpha + (1-z_i)\log(1-\alpha)\right) \\
&= -\left(\sum_{i=1}^{n} z_i\right)\log\alpha - \left(n - \sum_{i=1}^{n} z_i\right)\log(1-\alpha)
\end{aligned}$$

Differentiating and setting $\frac{d}{d\alpha}\left(-\log p(\mathcal{D}|\alpha)\right) = 0$, we find:

$$\begin{aligned}
\frac{d}{d\alpha}\left(-\log p(\mathcal{D}|\alpha)\right) &= -\frac{\sum_{i=1}^n z_i}{\alpha} + \frac{n - \sum_{i=1}^n z_i}{1-\alpha} = 0 \\
\implies \frac{\sum_{i=1}^n z_i}{\alpha} &= \frac{n - \sum_{i=1}^n z_i}{1-\alpha} \\
\implies (1-\alpha)\sum_{i=1}^n z_i &= \alpha(n - \sum_{i=1}^n z_i) \\
\implies \sum_{i=1}^n z_i - \alpha\sum_{i=1}^n z_i &= \alpha n - \alpha\sum_{i=1}^n z_i \\
\implies \sum_{i=1}^n z_i &= \alpha n \\
\implies \alpha &= \frac{1}{n}\sum_{i=1}^n z_i = \alpha_{\text{MLE}}
\end{aligned}$$

Question 24. [1 MARK]

Assume that $Y|X$ follows a Gaussian distribution with mean $\mu = xw_1$ and variance $\sigma^2 = \exp(xw_2)$ for all $x \in \mathbb{R}$ and $\mathbf{w} = (w_1, w_2)$ where $w_1, w_2 \in \mathbb{R}$. The negative log-likelihood, can be written as follows for a dataset $\mathcal{D} = ((x_1, y_1), \cdots, (x_n, y_n))$:

$$g(\mathbf{w}) = \sum_{i=1}^n g_i(\mathbf{w}) \quad \text{where } g_i(\mathbf{w}) = -\ln p(y_i|x_i, \mathbf{w}),$$

where $p(\cdot|\cdot)$ is the density of the above Gaussian distribution. What is partial derivative of g with respect to w_1?

a. $\frac{\partial g}{\partial w_1} = \sum_{i=1}^n \frac{x_i(y_i - x_i w_1)}{\exp(x_i w_2)}$

b. $\frac{\partial g}{\partial w_1} = \sum_{i=1}^n \frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)}$

c. $\frac{\partial g}{\partial w_1} = -\sum_{i=1}^n \frac{x_i(y_i - x_i w_1)}{\exp(x_i w_2)}$

d. $\frac{\partial g}{\partial w_1} = -\sum_{i=1}^n \frac{(y_i - x_i w_1)^2}{\exp(x_i w_2)}$

Solution:

Answer: c.
Explanation: To find $\frac{\partial g}{\partial w_1}$, note that the density of $Y|X$ is

$$p(y_i|x_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\exp(x_i w_2)}} \exp\left(-\frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)}\right)$$

The negative log-likelihood term $g_i(\mathbf{w})$ is:

$$g_i(\mathbf{w}) = \frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)} + \frac{1}{2}\ln(2\pi\exp(x_i w_2))$$

Differentiating $g_i(\mathbf{w})$ with respect to w_1 gives:

$$\frac{\partial g}{\partial w_1} = -\sum_{i=1}^{n} \frac{x_i(y_i - x_i w_1)}{\exp(x_i w_2)}\,.$$

Question 25. [1 MARK]

Let everything be defined as in the previous question. What is partial derivative of g with respect to w_2?

a. $\sum_{i=1}^{n} \left(-\frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)} + x_i\right)$

b. $\sum_{i=1}^{n} \left(\frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)} + \frac{x_i}{2}\right)$

c. $\sum_{i=1}^{n} \left(\frac{x_i(y_i - x_i w_1)^2}{2\exp(x_i w_2)} - \frac{x_i}{2}\right)$

d. $\sum_{i=1}^{n} \left(-\frac{x_i(y_i - x_i w_1)^2}{2\exp(x_i w_2)} + \frac{x_i}{2}\right)$

Solution:

Answer: d.
Explanation: To find $\frac{\partial g}{\partial w_2}$, we start with the expression for $g_i(\mathbf{w})$:

$$g_i(\mathbf{w}) = \frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)} + \frac{1}{2}\ln(2\pi \exp(x_i w_2))$$

Differentiating $g_i(\mathbf{w})$ with respect to w_2 gives:

$$\frac{\partial g}{\partial w_2} = \sum_{i=1}^{n} \left(-\frac{x_i(y_i - x_i w_1)^2}{2\exp(x_i w_2)} + \frac{x_i}{2}\right)\,.$$

Question 26. [1 MARK]

Let everything be defined as in the previous question. You want to solve for $\mathbf{w}_{\text{MLE}}$ using gradient descent. Using the partial derivatives you calculated in the previous quesitons, what would the gradient update rule look like with a constant step size η?

a. $w_1 \leftarrow w_1 - \eta \sum_{i=1}^{n} \left(\frac{x_i(y_i - x_i w_1)}{\exp(x_i w_2)}\right), \quad w_2 \leftarrow w_2 - \eta \sum_{i=1}^{n} \left(\frac{x_i(y_i - x_i w_1)^2}{2\exp(x_i w_2)} - \frac{x_i}{2}\right)$

b. $w_1 \leftarrow w_1 + \eta \sum_{i=1}^{n} \left(\frac{x_i(y_i - x_i w_1)}{\exp(x_i w_2)}\right), \quad w_2 \leftarrow w_2 + \eta \sum_{i=1}^{n} \left(\frac{x_i(y_i - x_i w_1)^2}{2\exp(x_i w_2)} - \frac{x_i}{2}\right)$

c. $w_1 \leftarrow w_1 - \eta \sum_{i=1}^{n} \left(\frac{(y_i - x_i w_1)}{2}\right), \quad w_2 \leftarrow w_2 - \eta \sum_{i=1}^{n} \left(\frac{(y_i - x_i w_1)^2}{2\exp(x_i w_2)} - \frac{x_i}{2}\right)$

d. $w_1 \leftarrow w_1 - \eta \sum_{i=1}^{n} \left(\frac{(y_i - x_i w_1)}{\exp(x_i w_2)}\right), \quad w_2 \leftarrow w_2 + \eta \sum_{i=1}^{n} \left(\frac{(y_i - x_i w_1)^2}{2} - \frac{x_i}{2}\right)$

Solution:

Answer: b.
Explanation: Plugging in the partial derivatives from the previous questions into the gradient descent update rule, we get:

$$w_1 \leftarrow w_1 + \eta \sum_{i=1}^{n} \frac{x_i(y_i - x_i w_1)}{\exp(x_i w_2)},$$

$$w_2 \leftarrow w_2 + \eta \sum_{i=1}^{n} \left(\frac{x_i(y_i - x_i w_1)^2}{2\exp(x_i w_2)} - \frac{x_i}{2} \right).$$