CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2024)
Quiz - 1

Date of Examination: 03.09.2024 Duration: 45 mins Total Marks: 15 marks

Instructions –

- Attempt all questions.
- MCQs have a single correct option.
- State any assumptions you have made clearly.
- Standard institute plagiarism policy holds.
- No evaluation without suitable justification.
- 0 marks if the option or justification of MCQs is incorrect.

1. What is the primary goal of using Maximum Likelihood Estimation (MLE) in linear regression? **[1 mark]**

   (A) To minimize the distance between the observed values and the model's predictions
   (B) To maximize the probability of observing the data given the parameters
   (C) To reduce the computational complexity of the model
   (D) To avoid overfitting the model to the training data

   Correct Answer: B, 1 mark for correct answer and correct reason
   Explanation: Maximum Likelihood Estimation (MLE) in linear regression is used to find the parameter values (weights) that maximize the likelihood of observing the given data. This approach focuses on maximizing the probability of the data given the model parameters. Reference: Lecture 3 - Linear Regression

2. In the context of bias and variance trade-off, what is the effect of increasing model complexity? **[1 mark]**

   (A) Decreases bias and variance
   (B) Increases bias and decreases variance
   (C) Decreases bias and increases variance
   (D) Increases both bias and variance

   Correct Answer: C, 1 mark for correct answer and correct reason
   Explanation: Increasing model complexity (e.g., using a higher-degree polynomial in regression) typically decreases bias, as the model can better capture the complexity of the underlying data pattern. However, it also increases variance as the model becomes more sensitive to fluctuations in the training data.

3. You are given a dataset of $n = 4$ independent and identically distributed (i.i.d.) observations x_1, x_2, x_3, x_4 drawn from a normal distribution with unknown mean μ and known variance $\sigma^2 = 1$. The observations are $x_1 = 2$, $x_2 = 3$, $x_3 = 2.5$, and $x_4 = 3.5$.

   The probability density function (PDF) of the normal distribution is:

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

   Which of the following is the Maximum Likelihood Estimate (MLE) of the mean μ based on this data? **[2 mark]** [Binary Marking- 1 mark for the correct answer and 1 mark for correctly deriving MLE equation]

   Options:

(A) $\mu = 2.5$
(B) $\mu = 3$
(C) $\mu = 2.75$
(D) $\mu = 3.25$

Explanation: To find the Maximum Likelihood Estimate (MLE) for μ, we first write down the likelihood function L(μ) for the given data. Since the variance

$$\sigma^2 = 1$$

is known, the likelihood function based on the normal distribution for the n = 4 observations is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}}$$

Taking the logarithm of the likelihood function to obtain the log-likelihood:

$$\ell(\mu) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To maximize the log-likelihood function, we differentiate it with respect to u and set the derivative equal to zero:

$$\frac{d\ell(\mu)}{d\mu} = \sum_{i=1}^{n}(x_i - \mu) = 0$$

Solving for μ:

$$u = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{4}(2 + 3 + 2.5 + 3.5) = 11/4 = 2.75$$

4. Describe the role of the cost function in logistic regression and how it differs from the cost function used in linear regression. Provide a formulaic representation. **[3 mark]** 1 mark for each part

   1. Role in Logistic Regression: The cost function in logistic regression, often a log-loss or cross-entropy loss, measures the discrepancy between the predicted probabilities and the actual class labels. It is designed to penalize wrong predictions with a high degree of certainty.
   2. Formulaic Representation: The cost function J(θ) for logistic regression is given by:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(h_\theta(x_i)) + (1 - y_i)\log(1 - h_\theta(x_i))]$$

where $(h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}})$ is the hypothesis function.

3. Difference from Linear Regression: Unlike linear regression's mean squared error, the logistic regression cost function is nonlinear and involves logarithmic terms, which better suits the binary nature of the dependent variable.

5. Derive the expression for the weights using the method of Least Squares in a simple linear regression model. Explain each step of your derivation. **[4 mark]** [1 mark for each part]

Expected Answer:

1. **Start with the Model Definition:** In simple linear regression, the model is defined as

$$y_i = w \cdot x_i + \epsilon_i$$

where y_i is the dependent variable, x_i is the independent variable, w is the weight or coefficient, and ϵ_i represents the error terms.

2. **Objective Function (Loss Function):** The Least Squares method aims to minimise the sum of the squares of the residuals (differences between observed and predicted values). The objective function is given by:

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

3. **Differentiation:** To find the value of w that minimises the loss function, differentiate L(w) with respect to w and set it to zero:

$$\frac{dL}{dw} = -2\sum_{i=1}^{n} x_i(y_i - w \cdot x_i) = 0$$

4. **Solve for w:** Rearranging the above equation:

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

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

This gives the Least Squares estimate of the weights in a simple linear regression model.

OR

Matrix Approach

$$E = \|y - Xw\|^2$$

where E is the sum of squared errors, y is the target vector, X is the data matrix and w the weights vector

The least-squares problem has an analytical solution. When differentiating the error by w, then finding w for when the derivative is equal to zero yields the pseudo-inverse solution:

$$\frac{\partial E}{\partial w} = 2X^T(y - Xw) = 0$$

$$X^T X w = X^T y$$

$$w = (X^T X)^{-1} X^T y$$

6. Suppose we have a data set with five predictors:

$X_1 =$ GPA,
$X_2 =$ IQ,
$X_3 =$ Level (1 for College and 0 for High School),
$X_4 =$ Interaction between GPA and IQ,
$X_5 =$ Interaction between GPA and Level.

The response variable is the starting salary after graduation (in thousands of dollars).

Suppose we use least squares to fit the model and obtain the following estimates for the coefficients:

$$\hat{\beta}_0 = 50,$$
$$\hat{\beta}_1 = 20,$$
$$\hat{\beta}_2 = 0.07,$$
$$\hat{\beta}_3 = 35,$$
$$\hat{\beta}_4 = 0.01,$$
$$\hat{\beta}_5 = -10.$$

Note: GPA greater than or equal to 3.5 is considered high on a scale of 4.

1. Which answer is correct, and why? [**2 marks**]

(A) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

(B) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

(C) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

(D) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

2. Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0. [**1 mark**]

3. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer. [**1 mark**]

Solution:

1. **Let:**

$$\begin{aligned}
x_1 &= \text{GPA} \\
x_2 &= \text{IQ} \\
x_3 &= \text{Level (College} = 1\text{, High School} = 0) \\
x_4 &= \text{Interaction between GPA and IQ } (x_1 \cdot x_2) \\
x_5 &= \text{Interaction between GPA and Level } (x_1 \cdot x_3)
\end{aligned}$$

Salary Equation:

$$\text{Salary} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 = 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_4 - 10x_5$$

For fixed IQ and GPA at x_1' and x_2':

$$\begin{aligned}
\text{Salary (high school)} &= 50 + 20x_1' + 0.07x_2' + 35 \times 0 + 0.01(x_1' \cdot x_2') - 10(x_1' \cdot 0) \\
&= 50 + 20x_1' + 0.07x_2' + 0.01(x_1' \cdot x_2') \\
\text{Salary (college)} &= 50 + 20x_1' + 0.07x_2' + 35 \times 1 + 0.01(x_1' \cdot x_2') - 10(x_1' \cdot 1) \\
&= 50 + 20x_1' + 0.07x_2' + 35 + 0.01(x_1' \cdot x_2') - 10x_1' \\
&= \text{Salary (high school)} + 35 - 10x_1'
\end{aligned}$$

From here:

$$\text{Salary (college)} - \text{Salary (high school)} = 35 - 10x_1'$$

Assuming the salary difference to be more than or equal to zero, we get:

$$35 - 10x_1' \geq 0 \quad \Rightarrow \quad x_1' \leq 3.5$$

Assuming the salary difference to be less than or equal to zero, we get:

$$35 - 10x_1' \leq 0 \quad \Rightarrow \quad x_1' \geq 3.5$$

Hence, for a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates, provided that the GPA is greater than or equal to 3.5. [Binary Marking- 1 mark for the correct answer and 1 mark for the correct reason with correct value]

2. Salary Calculation:

$$\text{Salary} = 50 + 20(4) + 0.07(110) + 35 + 0.01(110 \times 4) - 10(4)$$

$$\text{Salary} = 50 + 80 + 7.7 + 35 + 4.4 - 40 = 137.1 \text{ (in thousands of dollars)}$$

[Binary Marking]
For part1 and 2 marks will be awarded based on the correctness of the subsequent steps, regardless of the assumptions students make regarding the interaction. If they correctly apply their assumptions in the following steps, they deserve full marks.

3. False because the magnitude of the coefficient is not an indicator of statistical significance [Binary Marking]