Quiz3

1 MSE & MAE

(20pt) Consider the following true observed values and predicted values from a sample dataset:
Observed values: $y = [20, 21, 19, 30, 25, 100]$.
Predicted values: $\hat{y} = [22, 20, 18, 28, 24, 95]$.

(1) Calculate MSE and MAE of the given dataset.

Solution:

MSE: $\frac{1}{6} \cdot ((20-22)^2 + (21-20)^2 + (19-18)^2 + (30-28)^2 + (25-24)^2 + (100-95)^2) = 6$
MAE: $\frac{1}{6} \cdot (|20-22| + |21-20| + |19-18| + |30-28| + |25-24| + |100-95|) = 2$
Tips: Here are the definitions of MSE and MAE. However, in linear regression problems, we often introduce a factor of 1/2 to simplify the questions when calculating the gradient descent, as shown in Question 4. Thus, due to the potential ambiguity between Question 1 and Question 4, you only need to write down the calculation for the red-highlighted section and obtain the corresponding result to get full credit in this question.

(2) Discuss the impact of outliers on MSE and MAE.

MSE: MSE is more sensitive to outliers, which means even a single outlier can greatly increase the value of MSE.

MAE: MAE measures errors through absolute values, giving equal weight to all errors. Therefore, the value of MAE is not significantly affected by a single outlier, making MAE more robust.

2 Maximum Likelihood Estimate

(30pt) Suppose $X^1, X^2, \cdots, X^n$ are i.i.d discrete random variables, each following a Geometric distribution with parameter p. The probability mass function of the Geometric distribution is given by:

$$\Pr(X^i = x) = (1-p)^x p,$$

where p is an unknown parameter to estimate.

(1) What is the joint probability of observing the particular sample $\{x^1, \cdots, x^n\}$?

Solution:

$$\Pr((X^1 = x_1) \cap \cdots \cap (X^n = x_n)) = p^n \prod_{i=1}^{n} (1-p)^{x_i} = p^n (1-p)^{\sum_{i=1}^{n} x^i}$$

(2) Please explain the difference between the likelihood function and the joint probability distribution in this question.

Solution:

Joint probability: Given the parameter p, the probability of observing the data $X^1, X^2, \cdots, X^n$.

Likelihood function: Given the observed data $X_1, X_2, \cdots, X_n$. The likelihood function is used to estimate which parameter p makes the observed data most likely to occur.

(3) Write the **log likelihood function** and calculate the maximum likelihood estimate.

Solution:

$$\ln L_m(p; x^1, \cdots, x_n) = n \log p + (\sum_{i=1}^{n} x^i) \log (1-p)$$

$$\hat{p} = \frac{n}{n + \sum_{i=1}^{n} x^i}$$

3 Linear Regression

(30pt) True or False

1. In a linear regression model, all the parameters must be linear.
2. When the number of features is small, the normal equation can directly solve for the optimal parameters $\boldsymbol{\theta}$ of a linear regression problem without the need for iteration.

3. In the gradient descent method, the direction of parameter updates is the same as the direction of the gradient.

4. The pseudo-inverse matrix can be used to find the optimal parameters $\boldsymbol{\theta}$ in linear regression, even if the feature matrix $\mathbf{X}$ is not full-rank.

5. For linear regression, the least squares method is equivalent to maximum likelihood estimation when the errors follow a normal distribution.

6. Dummy coding uses only ones and zeros to convey all of the necessary information on category classification.

Solution:

FTFTTT

4 Gradient descent

(20pt) Given the data points $(1, 2), (2, 3), (3, 5)$, the loss function for linear regression is the mean squared error (MSE):

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

where $h_\theta(x_i) = \theta_0 + \theta_1 x_i$, and the number of data points is $m = 3$. The initial parameters are $\theta_0 = 0$, $\theta_1 = 0$, and the learning rate is $\alpha = 0.1$.

Using the gradient descent algorithm, compute one iteration of parameter updates, and provide the updated values for θ_0 and θ_1.

Solution:
Compute the partial derivatives:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) = \frac{1}{3} ((0 - 2) + (0 - 3) + (0 - 5)) = \frac{1}{3} \times (-10) = -\frac{10}{3}$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) x_i = \frac{1}{3} ((0 - 2) \cdot 1 + (0 - 3) \cdot 2 + (0 - 5) \cdot 3) = \frac{1}{3} \times (-23) = -\frac{23}{3}$$

Next, apply the gradient descent update rule with learning rate $\alpha = 0.1$:

$$\theta_0 := 0 - 0.1 \times \left(-\frac{10}{3}\right) = 0 + \frac{1}{3} \approx 0.333$$

$$\theta_1 := 0 - 0.1 \times \left(-\frac{23}{3}\right) = 0 + \frac{23}{30} \approx 0.767$$