



# Writing Assignment 1

Yushan Liu   Student ID: 2024214103

September 30, 2024

## Problem 1.1: Logistic Regression

### (a) Sigmoid Function

The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

We can rewrite the sigmoid function as:

$$\sigma(z) = \frac{e^z}{1 + e^z} = 1 - \frac{1}{1 + e^z}$$

Then the derivative of the sigmoid function with respect to  $z$  is:

$$\frac{d}{dz}\sigma(z) = \frac{d}{dz}\left(1 - \frac{1}{1 + e^z}\right) = \frac{e^z}{(1 + e^z)^2} = \frac{1}{1 + e^z} \cdot \frac{e^z}{1 + e^z} = \sigma(z) \cdot (1 - \sigma(z))$$

Thus, the derivative of the sigmoid function with respect to  $z$  is:

$$\frac{d}{dz}\sigma(z) = \sigma(z) \cdot (1 - \sigma(z))$$

### (b) Log-Likelihood Function

The log-likelihood function for logistic regression is given by:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^m (y^{(i)} \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})))$$

Since the log-likelihood is a sum over individual training examples, we can focus on the derivative for a single example  $i$ :

$$\ell_i(\boldsymbol{\theta}) = y^{(i)} \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}))$$

The derivative of the first term  $y^{(i)} \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})$  is:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}_j} (y^{(i)} \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) &= y^{(i)} \cdot \frac{1}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})} \cdot \sigma'(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}) \cdot x_j^{(i)} \\ &= y^{(i)} \cdot (1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) \cdot x_j^{(i)} \end{aligned}$$

The derivative of the second term  $(1 - y^{(i)}) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}))$  is:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}_j} ((1 - y^{(i)}) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}))) &= (1 - y^{(i)}) \cdot \frac{-1}{1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})} \cdot (-\sigma'(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) \cdot x_j^{(i)} \\ &= (1 - y^{(i)}) \cdot (-\sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) \cdot x_j^{(i)} \end{aligned}$$

Then combine the two terms and simplify, we have:

$$\frac{\partial}{\partial \boldsymbol{\theta}_j} \ell_i(\boldsymbol{\theta}) = (y^{(i)} - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) \cdot x_j^{(i)}$$

Summing over all training examples  $i = 1, \dots, m$ , we obtain the desired result:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^m (y^{(i)} - \sigma(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) x_j^{(i)}$$

## Problem 1.2: Ridge Regression

### (a) Gradient of the Ridge Regression Loss

We are given the ridge regression loss function:

$$J(\boldsymbol{\theta}) \triangleq \|\mathbf{y} - X\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

To compute the gradient with respect to  $\boldsymbol{\theta}$ , we first note that the loss function can be expanded as:

$$J(\boldsymbol{\theta}) = (\mathbf{y} - X\boldsymbol{\theta})^\top (\mathbf{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^\top \boldsymbol{\theta}$$

Now, differentiating  $J(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$ , we get:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -2X^\top (\mathbf{y} - X\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta}$$

Thus, the gradient is:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 2X^\top X\boldsymbol{\theta} - 2X^\top \mathbf{y} + 2\lambda\boldsymbol{\theta}$$

### (b) Gradient Descent Update Rule

Using the gradient computed above, the update rule for gradient descent is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t)$$

Substituting the gradient, we have:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha (2X^\top X\boldsymbol{\theta}_t - 2X^\top \mathbf{y} + 2\lambda\boldsymbol{\theta}_t)$$

### (c) Optimal Solution Using the Normal Equation

The optimal parameter  $\boldsymbol{\theta}^*$  can be derived by setting the gradient to zero:

$$2X^\top X\boldsymbol{\theta}^* - 2X^\top \mathbf{y} + 2\lambda\boldsymbol{\theta}^* = 0$$

Simplifying, we get the normal equation:

$$(X^\top X + \lambda I)\boldsymbol{\theta}^* = X^\top \mathbf{y}$$

Solving for  $\boldsymbol{\theta}^*$ , we obtain:

$$\boldsymbol{\theta}^* = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}$$

## Problem 1.3: Poisson Distribution and Generalized Linear Model (GLM)

### (a) Exponential Family Form of the Poisson Distribution

The probability mass function of the Poisson distribution is given by:

$$p(y | \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

We have known that a class of distributions is in the exponential family if its density can be written in the canonical form:

$$p(y | \eta) = h(y) \exp(\eta T(y) - a(\eta))$$

Rewriting in the exponential family form for the Poisson distribution, we have:

$$p(y | \eta) = \frac{1}{y!} \exp(\eta y - e^\eta)$$

We can see that:

- $\eta = \log(\lambda)$  is the natural parameter.
- $T(y) = y$  is the sufficient statistic.
- $a(\eta) = e^\eta$  is the log-partition function.
- $b(y) = \frac{1}{y!}$  normalizes the distribution.

### (b) GLM for Poisson Regression

From solving (a), we know that:

- $\eta = \log(\lambda)$  is the natural parameter.
- $T(y) = y$  is the sufficient statistic.

By deriving hypothesis function from the exponential family form, we have:

$$h_\theta(x) = E[T(y)|x; \theta] = \lambda = \eta$$

To adopt linear model  $\eta = \theta^T x$ , we have:

$$\log(\lambda) = \eta = \theta^T x$$

$$h_\theta(x) = e^{\theta^T x}$$

Thus, we can conclude that:

- **Hypothesis function:**  $h_\theta(x) = e^{\theta^T x}$ , where  $\theta^T x$  is the linear combination of the input features  $x$ .
- **Canonical link function:**  $g(\lambda) = \log(\lambda)$ , which relates the rate parameter  $\lambda$  to the natural parameter  $\eta = \theta^T x$ .
- **Inverse canonical link function:**  $g^{-1}(\eta) = e^\eta$ , which transforms the natural parameter  $\eta$  back into the rate parameter  $\lambda$ .

## Problem 1.4: Softmax Regression

The Softmax model's log-likelihood function is given by:

$$\ell(\Theta) \triangleq \sum_{i=1}^m \log p(y^{(i)} | x^{(i)}; \Theta) = \sum_{i=1}^m \sum_{l=1}^K \mathbf{1}\{y^{(i)} = l\} \log \frac{e^{\theta_l^\top x^{(i)}}}{\sum_{j=1}^K e^{\theta_j^\top x^{(i)}}}$$

We can express the log-likelihood function in terms of the indicator function and the softmax probabilities:

$$p(y^{(i)} = l | x^{(i)}; \Theta) = \frac{e^{\theta_l^\top x^{(i)}}}{\sum_{j=1}^K e^{\theta_j^\top x^{(i)}}}$$

The full log-likelihood can be written as:

$$\ell(\Theta) = \sum_{i=1}^m \log \left( \frac{e^{\theta_{y^{(i)}}^\top x^{(i)}}}{\sum_{j=1}^K e^{\theta_j^\top x^{(i)}}} \right) = \sum_{i=1}^m \left( \theta_{y^{(i)}}^\top x^{(i)} - \log \left( \sum_{j=1}^K e^{\theta_j^\top x^{(i)}} \right) \right)$$

What we need to calculate is:

$$\nabla_{\theta_l} \ell(\Theta) = \sum_{i=1}^m \frac{\partial}{\partial \theta_l} \left( \theta_{y^{(i)}}^\top x^{(i)} - \log \left( \sum_{j=1}^K e^{\theta_j^\top x^{(i)}} \right) \right)$$

We can take the derivative term-by-term:

1. Derivative of the first term  $\theta_{y^{(i)}}^\top x^{(i)}$

- For  $l = y^{(i)}$ ,  $\frac{\partial}{\partial \theta_l} \theta_{y^{(i)}}^\top x^{(i)} = x^{(i)}$ .
- For  $l \neq y^{(i)}$ ,  $\frac{\partial}{\partial \theta_l} \theta_{y^{(i)}}^\top x^{(i)} = 0$ .

2. Derivative of the second term  $-\log \left( \sum_{j=1}^K e^{\theta_j^\top x^{(i)}} \right)$

$$\begin{aligned} \frac{\partial}{\partial \theta_l} \left( -\log \left( \sum_{j=1}^K e^{\theta_j^\top x^{(i)}} \right) \right) &= -\frac{1}{\sum_{j=1}^K e^{\theta_j^\top x^{(i)}}} \cdot \frac{\partial}{\partial \theta_l} \left( \sum_{j=1}^K e^{\theta_j^\top x^{(i)}} \right) \\ &= -\frac{1}{\sum_{j=1}^K e^{\theta_j^\top x^{(i)}}} \cdot e^{\theta_l^\top x^{(i)}} \cdot x^{(i)} = -P(y = l | x^{(i)}; \Theta) \cdot x^{(i)} \end{aligned}$$

So for each class  $l$ , the gradient of the log-likelihood with respect to  $\theta_l$  is:

$$\nabla_{\theta_l} \ell(\Theta) = \sum_{i=1}^m (\mathbf{1}\{y^{(i)} = l\} - P(y = l | x^{(i)}; \Theta)) x^{(i)}$$

Where:

- $\mathbf{1}\{y^{(i)} = l\}$  is 1 if the  $i$ -th example belongs to class  $l$ , and 0 otherwise.
- $P(y = l | x^{(i)}; \Theta) = \frac{e^{\theta_l^\top x^{(i)}}}{\sum_{j=1}^K e^{\theta_j^\top x^{(i)}}}$  is the softmax probability.

## Problem 1.5: Maximun Likelihood Estimation

### (a) the Expression of Conditional Distribution

The conditional distribution of  $y$  given  $\mathbf{x}$  is the distribution of  $y - \boldsymbol{\theta}^\top \mathbf{x}$ , which is simply the distribution of the error term  $\epsilon$ . Hence, the conditional distribution of  $y$  given  $x$  is:

$$P_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{2\tau} \exp\left(-\frac{|y - \boldsymbol{\theta}^\top \mathbf{x}|}{\tau}\right)$$

### (b) the Log-Likelihood Function

Given the conditional probability  $P_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})$ , the log-likelihood for  $m$  samples  $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$  can be written as:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^m \log P_{Y|X}(y^{(i)}|x^{(i)}; \boldsymbol{\theta}) = \sum_{i=1}^m \log \left( \frac{1}{2\tau} \exp\left(-\frac{|y^{(i)} - \boldsymbol{\theta}^\top x^{(i)}|}{\tau}\right) \right) \\ &= \sum_{i=1}^m \left( \log\left(\frac{1}{2\tau}\right) - \frac{|y^{(i)} - \boldsymbol{\theta}^\top x^{(i)}|}{\tau} \right) \\ &= -m \log(2\tau) - \frac{1}{\tau} \sum_{i=1}^m |y^{(i)} - \boldsymbol{\theta}^\top x^{(i)}| \end{aligned}$$

### (c) the Geometric Interpretation of LAD

In ordinary least squares (OLS) regression, we minimize the sum of the squared distances between the predicted and actual values, effectively finding a line that minimizes the **squared Euclidean distance** between the points and the regression line. This gives the usual  $\ell_2$ -norm, which is sensitive to outliers because outliers have a disproportionately large influence due to the squaring of distances.

In least absolute deviation(LAD) regression, we minimize the sum of the absolute deviations  $|y^{(i)} - \boldsymbol{\theta}^\top x^{(i)}|$ , which corresponds to the  $\ell_1$ -norm.

The geometric interpretation of LAD is that instead of minimizing the Euclidean distance, we are minimizing the **Manhattan distance**, or the **vertical distances** between the data points and the regression line.

This results in a model that is more **robust to outliers** because outliers have a linear influence on the objective function, as opposed to a quadratic influence in OLS.