Written Assignment 1

Yushan Liu   Student ID: 2024214103

October 2, 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:

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

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

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

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

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

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

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

Then combine the two terms and simplify, we have:

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

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

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

Problem 1.2: Ridge Regression

(a) Gradient of the Ridge Regression Loss

We are given the ridge regression loss function:

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

To compute the gradient with respect to θ, we first note that the loss function can be expanded as:

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

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

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

Thus, the gradient is:

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

(b) Gradient Descent Update Rule

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

$$\theta_{t+1} :\ = \theta_t + \alpha\nabla_\theta J(\theta)$$

Substituting the gradient, we have:

$$\theta_{t+1} :\ = \theta_t + \alpha\left(2X^\top X\theta_t - 2X^\top\boldsymbol{y} + 2\lambda\theta_t\right)$$

(c) Optimal Solution Using the Normal Equation

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

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

Simplifying, we get the normal equation:

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

Solving for θ^*, we obtain:

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

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:

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

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

$$p(y \mid \eta) = h(y) \exp\left(\eta T(y) - a(\eta)\right)$$

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

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

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:

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

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

Thus, we can conclude that:

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

Problem 1.4: Softmax Regression

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

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

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

$$p(y^{(i)} = l \mid x^{(i)}; \mathbf{\Theta}) = \frac{e^{\boldsymbol{\theta}_l^\top x^{(i)}}}{\sum_{j=1}^{K} e^{\boldsymbol{\theta}_j^\top x^{(i)}}}$$

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:

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

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 $-\log \left(\sum_{j=1}^{K} e^{\theta_j^\top x^{(i)}} \right)$

$$\frac{\partial}{\partial \theta_l} \left(-\log \left(\sum_{j=1}^{K} e^{\theta_j^\top x^{(i)}} \right) \right) = -\frac{1}{\sum_{j=1}^{K} e^{\theta_j^\top x^{(i)}}} \cdot \frac{\partial}{\partial \theta_l} \left(\sum_{j=1}^{K} e^{\theta_j^\top x^{(i)}} \right)$$

$$= -\frac{1}{\sum_{j=1}^{K} e^{\theta_j^\top x^{(i)}}} \cdot e^{\theta_l^\top x^{(i)}} \cdot x^{(i)} = -P(y = l \mid x^{(i)}; \Theta) \cdot x^{(i)}$$

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

$$\nabla_{\boldsymbol{\theta}_l} \ell(\mathbf{\Theta}) = \sum_{i=1}^{m} \left(\mathbf{1}\{y^{(i)} = l\} - P(y = l \mid x^{(i)}; \Theta) \right) x^{(i)}$$

Where:

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

Problem 1.5: Maximun Likelihood Estimation

(a) the Expression of Conditional Distribution

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

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

(b) the Log-Likelihood Function

Given the conditional probability $P_{Y|X}(y|\boldsymbol{x};\boldsymbol{\theta})$, 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 ℓ_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 $|y^{(i)} - \theta^\top x^{(i)}|$, which corresponds to the ℓ_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.

References

[1] Andrew Ng, Tengyu Ma. *CCS 229 Lecture Notes*. Stanford University, 2023. Available online at: `https://cs229.stanford.edu/`

[2] Stephen Boyd, Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] OpenAI. *ChatGPT: A Conversational AI*. 2023. Available online at: `https://www.openai.com/chatgpt`

[4] K. L. Chung. *Stochastic Processes*. 2nd ed. Springer, 2001.