

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:

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

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

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

Then combine the two terms and simplify, we have:

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

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

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

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:

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

Substituting the gradient, we have:

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

(c) Optimal Solution Using the Normal Equation

The optimal parameter θ^* can be derived by setting the gradient to zero:

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

Simplifying, we get the normal equation:

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

Solving for θ^* , we obtain:

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

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 λ to the natural parameter $\eta = \theta^T x$.
- **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