Nguyen Vu Anh Ngoc
Machine Learning 1
October 5, 2023

Homework Week 6: Logistic Regression

Question 1. Sigmoid Function

a. Sigmoid function and its derivative.

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

Denote:

$$u(z) = e^{-z}, v(u) = 1+u, w(v) = \frac{1}{v}$$

The sigmoid function:

$$\sigma(z) = w(v(u(z)))$$

Chain rule:

$$\frac{d\sigma}{dz} = \frac{dw}{dv} \cdot \frac{dv}{du} \cdot \frac{du}{dz}$$

Calculate each part:

$$\frac{du}{dz} = \frac{d(e^{-z})}{dz} = -e^{-z}$$

$$\frac{dv}{du} = \frac{d(1+u)}{du} = 1$$

$$\frac{dw}{dv} = \frac{d(1/v)}{dv} = -\frac{1}{v^2} = \frac{dw}{dv} = -\frac{1}{(1+e^{-z})^2}$$

Multiplying the three derivatives together:

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

Using the identity:

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

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

The derivative can be expressed more compactly as:

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

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

b. Loss function

Hypothesis function:

$$\hat{y} = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

Likelihood function of a single observations

$$P(y^{(i)} | x^{(i)}; w) = (\hat{y}^{(i)})^{y^{(i)}} (1 - \hat{y}^{(i)})^{(1-y^{(i)})}$$

Likelihood function of the data set:

$$L(w) = \prod_{i=1}^{n} (\hat{y}^{(i)})^{y^{(i)}} (1 - \hat{y}^{(i)})^{(1-y^{(i)})}$$

Log-likelihood:

$$l(w) = \sum_{i=1}^{n} y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

Loss function (negative log-likelihood):

$$J(w) = -\sum_{i=1}^{n} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right]$$

This loss function is known as the **Binary Cross-Entropy Loss** or **Log Loss**. It is used for binary classification problems in logistic regression. The function quantifies the difference between the predicted probabilities ($\hat{y}$) by the model and the actual class labels.

c. Gradient vector for loss function

Given the loss function:

$$J(w) = -\frac{1}{n} \sum_{i=1}^{n} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right]$$

Where:

$$\hat{y}^{(i)} = \sigma(w^T x^{(i)}) = \frac{1}{1 + e^{-w^T x^{(i)}}}$$

The gradient of $J(w)$ with respect to w_j:

$$\frac{\partial J(w)}{\partial w_j} = \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)}$$

Compute the partial derivative with respect to each weight w_j:

$$\nabla J(w) = \begin{bmatrix} \frac{\partial J(w)}{\partial w_1} \\ \frac{\partial J(w)}{\partial w_2} \\ \vdots \\ \frac{\partial J(w)}{\partial w_n} \end{bmatrix} = \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)}) x^{(i)} = \mathbf{X}^T (\hat{\mathbf{y}} - \mathbf{y})$$

Partial derivative calculation:

1. Differentiate the loss with respect to $\hat{y}^{(i)}$:

$$\frac{\partial}{\partial \hat{y}^{(i)}} \left(-y^{(i)} \log(\hat{y}^{(i)}) - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right) = -\frac{y^{(i)}}{\hat{y}^{(i)}} + \frac{1 - y^{(i)}}{1 - \hat{y}^{(i)}}$$

2. Differentiate $\hat{y}^{(i)}$ (σ) w.r.t $w^T x^{(i)}$:

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

$$\frac{d\hat{y}^{(i)}}{dw^T x^{(i)}} = \hat{y}^{(i)} \cdot (1 - \hat{y}^{(i)})$$

3. Differentiate $w^T x^{(i)}$ with respect to w_j:

$$\frac{\partial (w^T x^{(i)})}{\partial w_j} = x_j^{(i)}$$

Combine all using **Chain rule**:

$$\begin{aligned}
\frac{\partial J(w)}{\partial w_j} &= \frac{d\sigma(z)}{dz} \cdot \frac{\partial z}{\partial w_j} \\
&= \sigma(z) \cdot (1 - \sigma(z)) \cdot x_j^{(i)} \\
&= \sum_{i=1}^{m} \left(-\frac{y^{(i)}}{\hat{y}^{(i)}} + \frac{1 - y^{(i)}}{1 - \hat{y}^{(i)}} \right) \cdot \hat{y}^{(i)} \cdot (1 - \hat{y}^{(i)}) \cdot x_j^{(i)} \\
&= \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)}
\end{aligned}$$

Question 2. Implement Logistic Regression

Question 3. MSE vs Negative Log-likelihood

Aspect	Binary Cross-Entropy (BCE)	Mean Squared Error (MSE)
Nature of Problem	Suited for binary classification with output range [0,1].	Naturally suited for regression problems with continuous and unbounded outputs.
Loss Surface	Convex, which means a single global minimum.	Can introduce non-convexities when used with logistic regression.
Interpretability	Directly models the negative log likelihood, making it probabilistically interpretable.	Less interpretable for probabilistic tasks; squared terms can disproportionately penalize outliers.
Outliers' Impact	Robust to outliers. An outlier with a very wrong prediction leads to a large but not disproportionately large loss.	The squaring function can lead to very large losses for outliers, causing a disproportionate impact.
Saturation & Gradients	Combined with the logistic sigmoid function, avoids saturation and the associated vanishing gradient problem.	Using sigmoid activations with MSE can lead to saturation, causing vanishing gradients.
Historical & Practical Use	Traditionally and widely used with logistic regression due to better empirical results.	Rarely used with logistic regression due to challenges like non-convexities and less robustness to outliers.

Table 1. Binary Cross-Entropy vs Mean Squared Error for Logistic Regression