Lecture 06: Logistic Regression

Yorick Chern

Foothill Machine Learning & Data Science Club

1 Introduction

For Linear Regression,

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

we used the mean-squared-error

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

as our cost function. The partial derivative is calculated in terms of θ and Gradient Descent is performed in order to find out the best model parameters θ. On the other hand, logistic regression is mainly used for classification tasks such as image recognition and email spam-ham filters. Through logistic regression, we would obtain a value known as the "probability," and this value will help us categorize where this instance belongs to, known as the class. For example: Setting 1 as spam and 0 as ham (the spam/ham class); if we input an email, and the model produces a probability of 0.75, then we can conclude that this email is most likely a spam email.

2 Logistic Regression

Just like a Linear Regression model, a Logistic Regression model computes a weighted sum of the input features (plus a bias term). However, instead of producing a direct prediction like Linear Regression does, it produces the Logistic of the result, a probability value $\hat{p}$ between 0 and 1. This is achieved by defining the sigma σ function where

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

with the following shape:

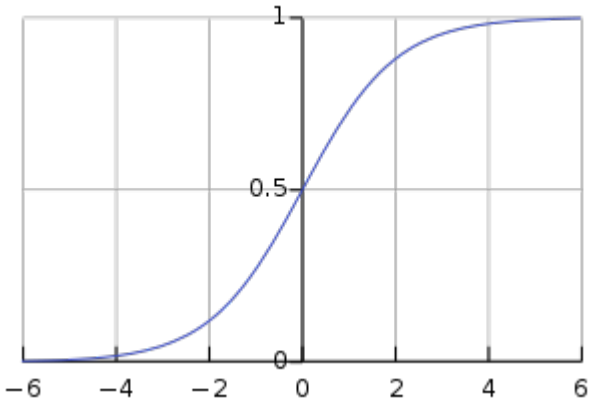

Figure 1: Shape of the Logistic Function

Once the Logistic Regression model has computed the probability

$$\hat{p} = h_\theta(x) = \sigma(\theta^T \cdot x)$$

it can make the prediction $\hat{y}$ using the piecewise function. Here, $\hat{p}$ is the probability of this instance belonging to the class $y = 1$.

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} > 0.5 \end{cases}$$

1. $\hat{y}$ is the predicted class
2. When $\hat{y} = 0$, $\hat{p} < 0.5$, $t < 0$
3. $\hat{y} = 1$, $\hat{p} > 0.5$, $t > 0$
4. $\hat{y}$ could represent different classes where, for example, 1 is spam and 0 is ham emails

3 Cost Function

Just like Linear Regression, Logistic Regression also needs a cost function in order to evaluate whether the model is becoming more accurate. This is the cost function for a single training instance:

$$cost(\theta) = \begin{cases} -\log(\hat{p}) & \text{if } y = 1 \\ -\log(1 - \hat{p}) & \text{if } y = 0 \end{cases}$$

Note: y is the actual class this training instance belongs to. This piecewise makes perfect sense because if the instance belongs to $y = 1$, then the higher

the probability the lower the penalty. For example: $-\log(0.98) = 0.0088$. Conversely, if the instance actually belongs to $y = 0$, then the lower the probability the lower the penalty. For example: With a probability of 0.012, meaning that the chance of this instance belonging to $y = 1$ is 0.012 when in reality it belongs to the class $y = 0$, produces a low penalty of $-\log(1 - 0.012) = 0.0052$.
Now, we simply need to sum all the penalty of each instance. This is the Logistic Regression Cost Function:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$$

Then, the same procedure is done. The partial derivative with respect to θ is used for the same Gradient Descent algorithm discussed in the Linear Regression lecture.

4 Softmax Regression

What happens if our classification is more than binary (e.g. spam/ham)? For example, we might need to classify images with handwritten digits from 0 to 9 on them.
First, when given an instance x, the Softmax Regression computes a score $s_k(x)$ for each class k

$$s_k(x) = x^T \theta_k$$

Softmax Regression simply computes the probability $\hat{p}_k$ that the instance belongs to class k by running the scores through the softmax function:

$$\hat{p}_k = \frac{\mathrm{e}^{\mathrm{s_k(x)}}}{\sum_{j=1}^{K} \mathrm{e}^{\mathrm{s_j(x)}}}$$

Then, the Softmax Regression classifier predicts the class with the highest estimated probability (which is the class with the highest score).

5 Conclusion

Obviously, this chapter only covers the tip of the iceberg of Logistic Regression and Softmax Regression. I wholeheartedly suggest anyone interested in Classification tasks to read more about this topic as it is extremely interesting and beneficial!