Linear Regression

Hanxiao Du

Contents

1 Introduction

Although linear regression can be used in classification tasks, it usually performs poorly. For example, instances that are far from the centroid could drag the regression line so that the linear regression cannot yield a proper dicision boundary (see figure 1).

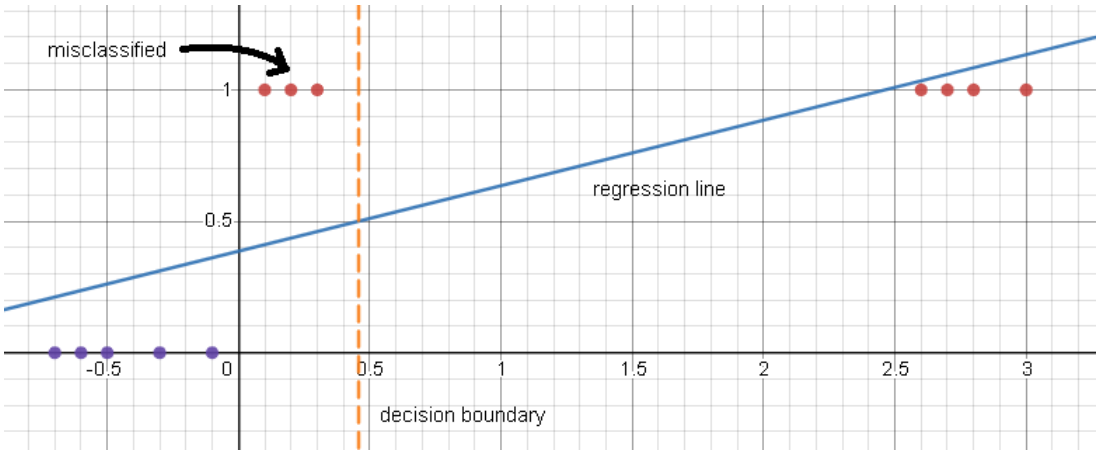

Figure 1: Linear regression as a classifier.

Logistic regression can solve this issue by fitting the data with logistic (sigmoid) function (see figure 2):

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

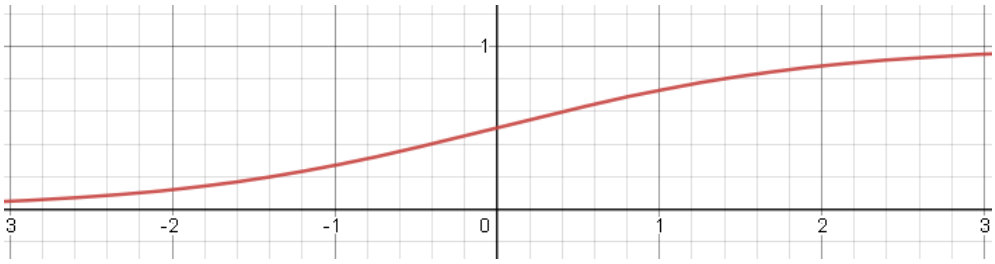

Figure 2: Logistic (sigmoid) function.

2 Hypothesis

Logistic regression assumes the real binary outcome random variable $Y = \mathbb{I}(X\beta + \epsilon \geq 0)$, where $\mathbb{I}(\cdot)$ is the indicator function, X is the independent random variable, β is the parameter (coefficent) vector and ϵ is the error term matrix. The distribution of the error term ϵ conditional on $X = x$ is $\epsilon|X = x \sim Logistic(\mu = 0, s = 1)$.
The logistic distribution $Logistic(\mu = 0, s = 1)$ has the sigmoid function as its cumulative distribution function(CDF):

$$F(z) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Thus, the probability of the outcome variable $Y = 1$ given X can be calculated as:

$$\begin{aligned}
\mathbb{P}(Y = 1|X = x) &= \mathbb{P}(X\beta + \epsilon \geq 0|X = x) && (3)\\
&= \mathbb{P}(\epsilon \geq -X\beta|X = x) && (4)\\
&= 1 - \mathbb{P}(\epsilon < -X\beta|X = x) && (5)\\
&= 1 - F(-x^T\beta), \text{ where } F \text{ is the CDF defined in (2)} && (6)\\
&= 1 - \sigma(-x^T\beta) && (7)\\
&= 1 - (1 - \sigma(x^T\beta)), \text{ by (1)} && (8)\\
&= \frac{1}{1 + e^{-x^T\beta}} && (9)
\end{aligned}$$

Then, we can model $\mathbb{P}(Y = 1|X = x)$ by $h_\beta(x) = \sigma(x^T\beta) = \frac{1}{1+e^{-x^T\beta}}$
Additionally,

$$\begin{aligned}
\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + e^{-x^T\beta}} &\iff \frac{1 - \mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 1|X = x)} = e^{-x^T\beta} && (10)\\
&\iff \frac{\mathbb{P}(Y = 1|X = x)}{1 - \mathbb{P}(Y = 1|X - x)} = e^{x^T\beta} && (11)\\
&\iff log(\frac{\mathbb{P}(Y = 1|X = x)}{1 - \mathbb{P}(Y = 1|X = x)}) = x^T\beta && (12)\\
&\iff logit(\mathbb{P}(Y = 1|X = x)) = x^T\beta && (13)
\end{aligned}$$

where

$$logit(p) = log(\frac{p}{1 - p}) = \sigma^{-1}(p) \tag{14}$$

3 Decision Boundary

The decision boundary of logistic regression is $\mathbb{P}(Y = 1|X = x) = 0.5$, since when $\mathbb{P}(Y = 1|X = x) > 0.5$, the model suggests that $\mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x)$ and vice versa.

Or equivalently,

$$\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + e^{-x^T\beta}} = 0.5 \iff x^T\beta = 0 \tag{15}$$

4 Cost Function

The loss function for logistic regression is normally the binary cross entropy, thus we have the cost:

$$J(\beta) = H_\beta(\underset{\sim}{y}, \underset{\sim}{\hat{y}}) = -||\underset{\sim}{y}log(h_\beta(\underset{\sim}{x})) + (\underset{\sim}{1} - \underset{\sim}{y})log(1 - h_\beta(\underset{\sim}{x}))||^2 \tag{16}$$

$$= -\sum_{i=1}^{n}\Big[y_i log(h_\beta(x_i)) + (1 - y_i)log(1 - h_\beta(x_i))\Big] \tag{17}$$

$$= -\sum_{i=1}^{n}\Big[y_i log(\frac{1}{1 + e^{-x_i^T\beta}}) + (1 - y_i)log(1 - \frac{1}{1 + e^{-x_i^T\beta}})\Big] \tag{18}$$

$$= -\sum_{i=1}^{n}\Big[y_i log(\frac{1}{1 + e^{-x_i^T\beta}}) + (1 - y_i)log(\frac{e^{-x_i^T\beta}}{1 + e^{-x_i^T\beta}})\Big] \tag{19}$$

The reason why we do not use the mean squared error as the cost function is that we cannot guarantee the convexity of the mean squared error when it is applied to logistic regression.

Theorem 1. *$J(\beta)$ is convex.*

Proof.

$$\frac{\partial J(\beta)}{\partial \beta} = -\sum_{i=1}^{n}\Big[y_i\frac{e^{-x_i^T\beta}}{1 + e^{-x_i^T\beta}}x_i - (1 - y_i)\frac{1}{1 + e^{-x_i^T\beta}}x_i\Big] \tag{20}$$

$$= -\sum_{i=1}^{n}\Big[y_i(1 - h_\beta(x_i))x_i - (1 - y_i)h_\beta(x_i)x_i\Big] \tag{21}$$

$$= -\sum_{i=1}^{n}\Big[y_i - y_i h_\beta(x_i) - h_\beta(x_i) + y_i h_\beta(x_i)\Big]x_i \tag{22}$$

$$= -\sum_{i=1}^{n}\Big[y_i - h_\beta(x_i)\Big]x_i \tag{23}$$

$$= -\sum_{i=1}^{n}\Big[y_i - \sigma(x_i^T\beta)\Big]x_i \tag{24}$$

$$= \sum_{i=1}^{n}\Big[\sigma(x_i^T\beta) - y_i\Big]x_i \tag{25}$$

$$= \sum_{i=1}^{n}\Big[\frac{1}{1 + e^{-x_i^T\beta}} - y_i\Big]x_i \tag{26}$$

$$\frac{\partial^2 J(\beta)}{\partial \beta^2} = \frac{\partial}{\partial \beta}\Big(\frac{\partial J(\beta)}{\partial \beta}\Big) = \sum_{i=1}^{n} \frac{e^{-x_i^T \beta}}{(1+e^{-x_i^T \beta})^2} x_i x_i^T \tag{27}$$

where $x_i x_i^T$ is always positive semi-definite.
Thus, $\frac{\partial^2 J(\beta)}{\partial \beta^2}$ is always positive semi-definite, which implies that $J(\beta)$ is convex.

■

There is no analytical solution for $\hat{\beta} = \underset{\beta}{\text{argmin}} J(\beta)$. Gradient descent technique can be used to find the numerical approximation of $\hat{\beta}$:

$$\beta_{t+1} \leftarrow \beta_t - \eta \frac{\partial J(\beta)}{\partial \beta} = \beta_t - \eta \sum_{i=1}^{n} \Big[\sigma(x_i^T \beta) - y_i\Big] x_i \tag{28}$$