Logistic Regression & Classification

Md. Mohsin Uddin

East West University

mmuddin@ewubd.edu

May 20, 2019

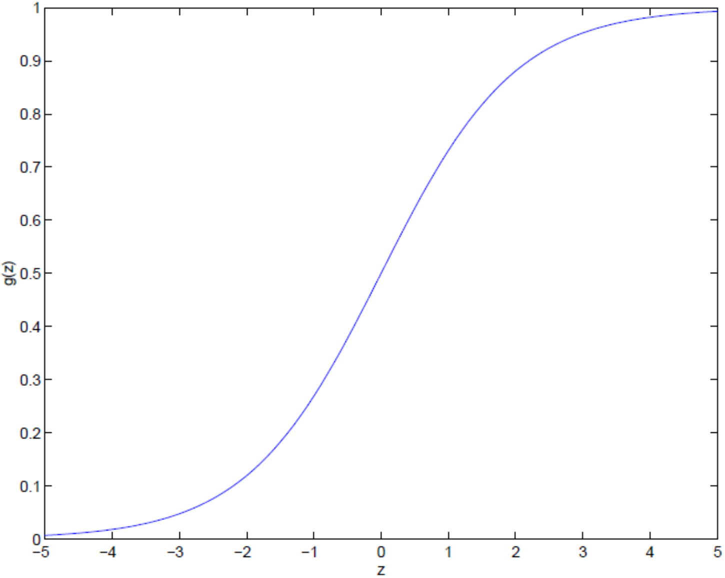
Classification

- Classification predicts only a small number of discrete values.
- Binary classification problem predicts only two values, 0 and 1.
- For instance, if we are trying to build a spam classifier for email, then $x^{(i)}$ may be some features of a piece of email, and y may be 1 if it is a piece of spam mail, and 0 otherwise. 0 is also called the negative class, and 1 the positive class.
- Given $x^{(i)}$, the corresponding $y^{(i)}$ is also called the label for the training example.

Logistic Regression (Intuition)

- Consider, the previous Linear Regression hypothesis.
- Intuitively, it also doesn't make sense for $h_\theta(x)$ to take values larger than 1 or smaller than 0 when we know that $y \in 0, 1$.
- Logistic regression hypothesis, $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$
  Here, $g(z) = \frac{1}{1+e^{-z}}$ is called the Logistic function or Sigmoid function.

Logistic function plot



Derivative of Logistic/Sigmoid function

$$
\begin{aligned}
g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\
&= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\
&= g(z)(1-g(z)).
\end{aligned}
$$

Logistic Regression : Problem definition

Let's assume,
$P(y = 1|x; \theta) = h_\theta(x)$
$P(y = 0|x; \theta) = 1 - h_\theta(x)$

This can be written more compactly as,
$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$

Logistic Regression : Problem definition

Assuming that the m training examples were generated independently, we can then write down the likelihood of the parameters as

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

As before, it will be easier to maximize the log likelihood:

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned}$$

Logistic Regression : Gradient Ascent rule for maximization

Repeatedly perform the following update:

$$\theta := \theta + \alpha \nabla_\theta \ell(\theta)$$

Logistic Regression : Stochastic Gradient Ascent rule

Consider, only one training example (x, y) and take derivative to derive the stochastic gradient ascent rule

$$\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial \theta_j}g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)g(\theta^T x)(1-g(\theta^T x))\frac{\partial}{\partial \theta_j}\theta^T x \\
&= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right)x_j \\
&= \left(y - h_\theta(x)\right)x_j
\end{aligned}$$

This therefore give us the stochastic gradient ascent rule:

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

References

- Christopher M. Bishop, Pattern recognition and Machine learning. Springer, 2006.
- Tom Mitchell, Machine learning. McGraw-Hill, 1997
- Lecture Notes of Andrew Ng