Foundations of Machine Learning

Module 5:
Part A: Logistic Regression

Sudeshna Sarkar
IIT Kharagpur

Logistic Regression for classification

- Linear Regression:

$$h(x) = \sum_{i=0}^{n} \beta_i x_i = \beta^T X$$

- Logistic Regression for classification:

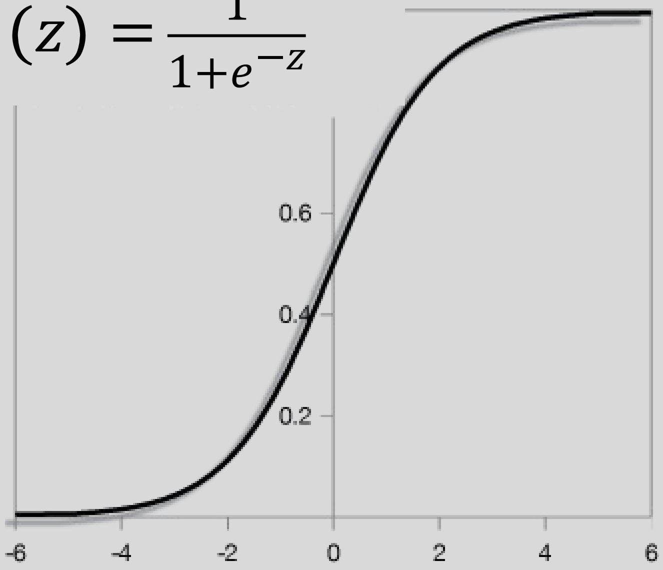
$$h_\beta(x) = \frac{1}{1+e^{-\beta^T X}} = g(\beta^T x)$$

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

is called the logistic function or the sigmoid function.

Logistic:

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



Sigmoid function properties

- Bounded between 0 and 1
- $g(z) \to 1$ as $z \to \infty$
- $g(z) \to 0$ as $z \to -\infty$

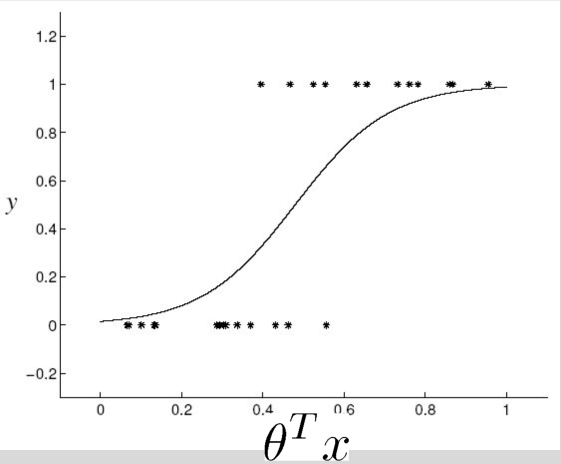
$g'(z) = \frac{d}{dz}\frac{1}{1+e^{-z}}$

$= \frac{1}{(1+e^{-z})^2}.e^{-z}$

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

$= g(z)(1-g(z)$

Logistic regression (sigmoid classifier)



- The condition distribution: a Bernoulli

$$p(y \mid x) = h(x)^y (1 - h(x))^{1-y}$$

where

$$h(x) = \frac{1}{1 + e^{-\beta^T x}}$$

- We can used the gradient method as in Linear Regression

Logistic Regression

- In logistic regression, we learn the conditional distribution P(y|x)
- Let $p_y(x; \beta)$ be our estimate of P(y|x), where β is a vector of adjustable parameters.
- Assume there are two classes, y = 0 and y = 1 and

$$P(y=1|x) = h_\beta(x)$$

$$P(y=0|x) = 1 - h_\beta(x)$$

- Can be written more compactly

$$P(y|x) = h(x)^y (1 - h(x))^{1-y}$$

- We can used the gradient method

Maximize likelihood

$$L(\beta) = p(\vec{y}|X;\beta)$$
$$= \prod_{i=1}^{m} p(y_i|x_i;\beta)$$
$$= \prod_{i=1}^{m} h(x_i \ )^{y_i}(1 - h(x_i))^{1-y_i}$$
$$l(\beta) = \log(L(\beta))$$
$$= \sum_{i=1}^{m} y^i \log h(x^i) + (1 - y_i)(\log(1 - h(x_i))$$

$l(\beta) = \sum_{i=1}^{m} y^i \log h(x^i) + (1-y_i)(\log(1-h(x_i))$

- How do we maximize the likelihood? Gradient ascent
  - Updates: $\beta = \beta + \alpha \nabla_\beta l(\beta)$

  Assume one training example (x,y), and take derivatives to derive the stochastic gradient ascent rule.

$$\frac{\partial}{\partial \beta_j} l(\beta)$$

$$= \left(\left(y \frac{1}{g(\beta^T(x)} \right) - (1-y) \frac{1}{1-g\left(\beta^T x\right)} \right) \frac{\partial}{\partial \beta_j} g(\beta^T x)$$

$$= \left(\left(y \frac{1}{g(\beta^T(x)} \right) - (1-y) \frac{1}{1-g\left(\beta^T x\right)} \right) g(\beta^T x)(1-g(\beta^T x) \frac{\partial}{\partial \beta_j} \beta^T x$$

$$= (y\left(1-g\left(\beta^T x\right)\right) - (1-y)\, g\left(\beta^T x\right))x_j$$

$$= (y - h_\beta(x))x_j$$

$$\beta = \beta + \alpha \nabla_{\beta} l(\beta)$$

$$\beta_j = \beta_j + \alpha(y^{(i)} - h_{\beta}\left(x^{i}\right))x_j^{(i)}$$