

## Logistic regression

(a) Given that

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \log(1 + e^{-y^{(i)} \theta^T x^{(i)}})$$

we can get

$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{m} \sum_{k=1}^m \frac{-y^{(k)} x_i^{(k)}}{1 + e^{y^{(k)} \theta^T x^{(k)}}}$$

then

$$\frac{\partial J(\theta)}{\partial \theta_i \partial \theta_j} = \frac{1}{m} \sum_{k=1}^m \frac{x_i^{(k)} x_j^{(k)} e^{y^{(k)} \theta^T x^{(k)}}}{(1 + e^{y^{(k)} \theta^T x^{(k)}})^2}$$

which is  $H_{ij}$

so

$$\begin{aligned} Z^T H Z &= \sum_{i=1}^n \sum_{j=1}^n z_i H_{ij} z_j \\ &= \frac{1}{m} \sum_{k=1}^m \frac{\sum_{i=1}^n \sum_{j=1}^n z_i x_i^{(k)} x_j^{(k)} z_j e^{y^{(k)} \theta^T x^{(k)}}}{(1 + e^{y^{(k)} \theta^T x^{(k)}})^2} \end{aligned}$$

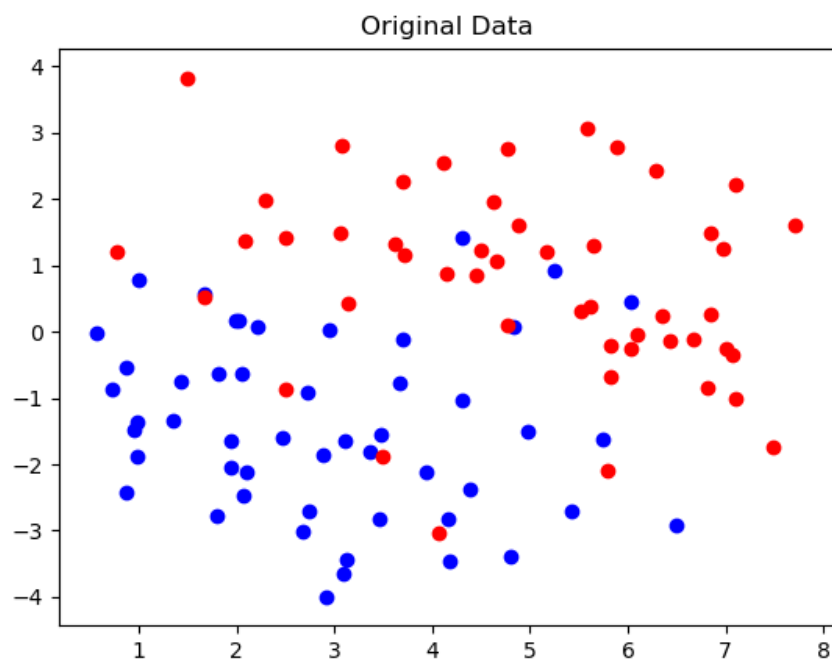
known that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n z_i x_i^{(k)} x_j^{(k)} z_j &= (X^T Z)^2 \geq 0 \\ \frac{e^{y^{(k)} \theta^T x^{(k)}}}{(1 + e^{y^{(k)} \theta^T x^{(k)}})^2} &> 0 \end{aligned}$$

we can easily get

$$Z^T H Z \geq 0$$

(b) Firstly, we plot the original data



To implement Newton's Method, we calculate the value of

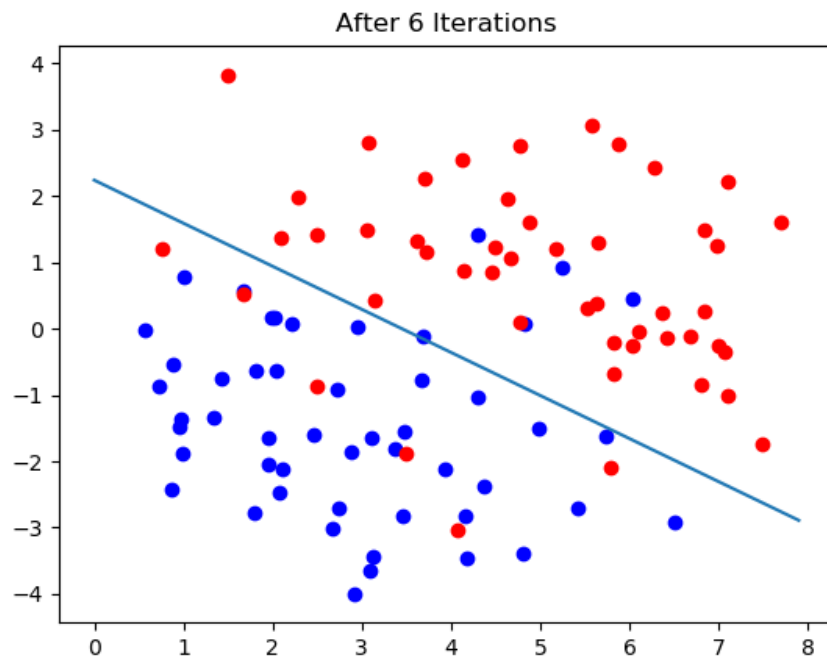
$$\frac{1}{m} \sum_{k=1}^m \frac{-y^{(k)} x_i^{(k)}}{1 + e^{y^{(k)} \theta^T x^{(k)}}}$$

and **Hessian**

then using the **update rule**

$$\theta := \theta - H^{-1} \nabla_{\theta} l(\theta)$$

(c) Through 6 iterations, we finally get the result



and the  $\theta$  is  $[-2.6205116, 0.76037154, 1.17194674]$ .

# Poisson regression and the exponential family

(a) Firstly, we get

$$\begin{aligned} p(y; \lambda) &= \frac{e^{-\lambda} \lambda^y}{y!} \\ &= \frac{1}{y!} \exp(\log \lambda^y - \log e^\lambda) \\ &= \frac{1}{y!} \exp(\log \lambda y - \lambda) \end{aligned}$$

It's easy to know that

$$\begin{aligned} \log \lambda &= \eta, \lambda = e^\eta \\ T(y) &= y \\ a(\eta) &= -e^\eta \\ b(y) &= \frac{1}{y!} \end{aligned}$$

and Poisson distribution is in the exponential family.

(b) Known that if  $Y \sim P(\lambda)$ , then  $E[Y] = \lambda$ ,

$$\begin{aligned} h_\theta(x) &= E[y|x; \theta] \\ &= \lambda \\ &= e^\eta = e^{\theta^T x} \end{aligned}$$

so the response function for the family is

$$g(\eta) = e^\eta$$

.

(c) Firstly, we can get likelihood function  $L(\theta)$

$$\begin{aligned} L(\theta) &= \prod_{k=1}^m P(y^{(k)}|x^{(k)}; \lambda) \\ &= \prod_{k=1}^m \frac{e^{-\lambda} \lambda^{y^{(k)}}}{y^{(k)}!} \end{aligned}$$

Then, the log-likelihood function  $l(\theta)$

$$\begin{aligned} l(\theta) &= \sum_{k=1}^m (-\log y^{(k)}!) + (-\lambda) + y^{(k)} \log \lambda \\ &= \sum_{k=1}^m (-\log y^{(k)}!) + (-e^{\theta^T x^{(k)}}) + y^{(k)} \theta^T x^{(k)} \end{aligned}$$

Then, we can get the derivative of the log-likelihood with respect to  $\theta_i$

$$\frac{\partial l(\theta)}{\partial \theta_i} = \sum_{k=1}^m (y^{(k)} - e^{\theta^T x^{(k)}}) x_i^{(k)}$$

So, the stochastic gradient ascent learning rule is

$$\theta_i = \theta_i + \alpha (y^{(k)} - e^{\theta^T x^{(k)}}) x_i^{(k)}$$

(d) Firstly, we can know that

$$p(y|x; \theta) = b(y) \exp(\eta^T y - a(\eta))$$

then, for a simple training data  $(x, y)$  in stochastic gradient ascent,

$$\frac{\partial p(y|x; \theta)}{\partial \theta_i} = x_i (y - \frac{\partial a(\eta)}{\partial \eta})$$

known that,

$$\int_y p(y|x; \eta) = 1$$

we can imply derivation in both sides, and get

$$\int_y p(y|x; \eta) (y - \frac{\partial a(\eta)}{\partial \eta}) = 0$$

finally, we get

$$\frac{\partial a(\eta)}{\partial \eta} = \int_y p(y|x; \eta) y = h_\theta(x)$$

so the gradient we get is

$$\frac{\partial p(y|x; \theta)}{\partial \theta_i} = x_i (y - h_\theta(x))$$

the update rule is

$$\theta_i = \theta_i - \alpha (h_\theta(x) - y) x_i$$

# Gaussian discriminant analysis

(a) gg