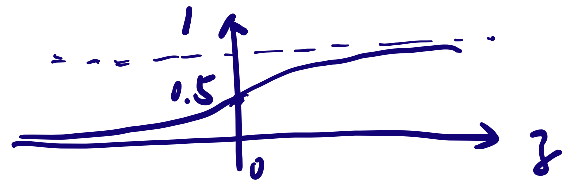
Sep 26, 2022 Mon

$p(y|x) \sim \text{Bernoulli}(\theta)$

$\theta = \text{sigmoid}(w^T x)$

Sigmoid function

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




$p(y|x) \sim \text{Ber}(\theta)$

$= \text{Ber}(\sigma(z))$

$= \text{Ber}\left(\frac{1}{1+\exp(-w^T x)}\right)$

Training of Logistic Regression

Goal: estimate $w$, Given $\{x_i, y_i\}$

Testing of LR

Given a new $x^*$.

$\theta = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-\hat{w}^T x^*)}$

if $\theta > 0.5$, then $y^* = 1$

else $y^* = 0$

$P(y^*=1) = \theta_i = \frac{1}{1+\exp(-w^T x_i^*)}$

$P(y^*=0) = 1-\theta_i$

# Learning Parameters

MLE approach.

Likelihood $P(D|\theta) = \prod_{i=1}^{N} \theta_i^{y_i} (1-\theta_i)^{1-y_i}$

$$NLL = -\log P(D|\theta)$$
$$= \sum_{i=1}^{N} -y_i \log \theta_i - (1-y_i)\log(1-\theta_i)$$

Cross-entropy loss $H(p,q) = -\sum_k p \log(q)$

log-loss

MLE = min NLL

Gradient Descent

$$W_{k+1} = W_k - \eta \frac{dNLL}{dw}$$

$$\theta_i = \sigma(z) = \frac{1}{1+\exp(-w^T x)}$$

$$\frac{d \log \theta_i}{w} = \frac{1}{\theta_i} \cdot \frac{d\theta_i}{dw} = \frac{1}{\theta_i} \cdot \frac{d\sigma}{dw}$$

$$\frac{d\sigma}{dw} = \cancel{-}(1+\exp(-w^T x))^{-2} \cdot \exp(-w^T x) \cdot (\cancel{-}X)$$
$$= \frac{\exp(-w^T x)}{(1+\exp(-w^T x))^2} \cdot X$$
$$= \sigma(1-\sigma) \cdot X$$

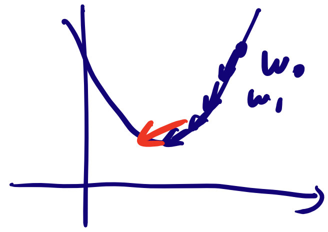
$$\frac{dNLL}{dw}$$

$$= \sum_{i=1}^{N} -y_i \frac{1}{\cancel{\theta_i}} \cancel{\theta_i}(1-\theta_i) \cdot x_i$$
$$+ (1-y_i) \frac{1}{\cancel{1-\theta_i}} \cancel{(1-\theta_i)} \theta_i x_i$$
$$= \sum_{i=1}^{N} (\cancel{y_i \theta_i x_i} - y_i x_i + \theta_i x_i - \cancel{y_i \theta_i x_i})$$
$$= \sum_{i=1}^{N} (\theta_i - y_i) x_i$$

$$w_{k+1} = w_k - \eta \frac{dNLL}{dw}$$

$\eta$ learning rate

$$\frac{dNLL}{dw} = \sum_{i=1}^{N} (\theta_i - y_i) x_i$$




Newton's method

$$w_{k+1} = w_k - H_k^{-1} \frac{dNLL}{dw_k}$$

$H$ : Hessian matrix

Second order derivative of a function.

$$f(x) \qquad f'(x) = \frac{df(x)}{dx} \qquad f''(x) = \frac{d^2 f(x)}{dx^2}$$

$$f(x) = x^3 + 2x^2$$

$$f'(x) = 3x^2 + 4x$$

$$f''(x) = 6x + 4$$

$$f(w) \qquad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}_{d \times 1}$$

$$\frac{df(w)}{dw} = \begin{bmatrix} \frac{df(w)}{dw_1} \\ \vdots \\ \frac{df(w)}{dw_d} \end{bmatrix}_{d \times 1}$$

$$H = \frac{d^2 f(w)}{dw^2} = \begin{bmatrix} \frac{d^2 f(w)}{dw_1^2} & \cdots & \frac{d^2 f(w)}{dw_1 dw_d} \\ \vdots & & \\ \frac{d^2 f(w)}{dw_d dw_1} & \cdots & \frac{d^2 f(w)}{dw_d^2} \end{bmatrix}_{d \times d}$$

$$f(w) = w_1 w_2 + 2w_1^2 + 3w_1^2 w_2 + 4$$

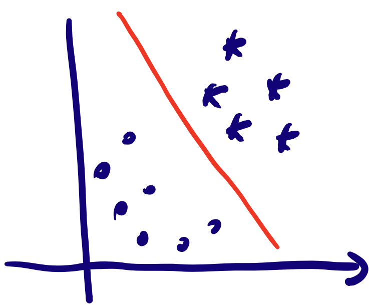
$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}_{2\times 1}$$

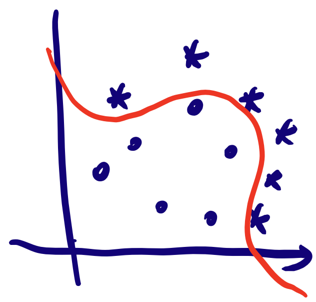
$$\frac{df(w)}{dw} = \begin{bmatrix} w_2 + 4w_1 + 6w_1 w_2 \\ w_1 + 3w_1^2 \end{bmatrix}_{2\times 1}$$

$$H = \begin{bmatrix} 4 + 6w_2 & 1 + 6w_1 \\ 1 + 6w_1 & 0 \end{bmatrix}_{2\times 2}$$

For logistic Regression.

$$H = -\sum_{i=1}^{N} \theta_i (1-\theta_i) x_i x_i^T$$





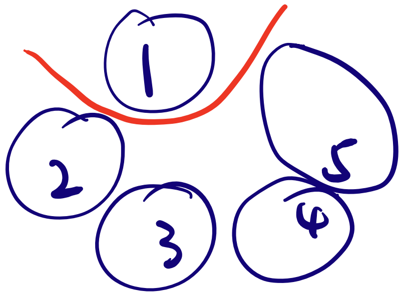
$$\phi = [1, x, x^2 \cdots x^d]$$

$$\theta_i = \sigma(z) = \frac{1}{1+\exp(-w^T \phi(x))}$$

$$L(w) = NLL(w) + \frac{1}{2}\lambda w^T w$$

$$\frac{dL(w)}{dw} = \frac{dNLL}{dw} + \lambda w$$

$$\hat{H} = H + \lambda I$$




Multiple classes, multi class / multi label

$P(y|x) \sim \text{Multinoulli}(\theta)$ $\theta = [\theta_1 \cdots \theta_C]$

$$\theta_j = \frac{\exp(w_j^T x)}{\sum_{k=1}^{C} \exp(w_k^T x)}$$ softmax function

$C = 2$

$$\theta_1 = \frac{\exp(w_1^T x)}{\exp(w_1^T x) + \exp(w_2^T x)}$$

$$= \frac{1}{1 + \exp((w_2 - w_1)^T x)}$$

$$= \frac{1}{1 + \exp(-\hat{w}^T x)} \qquad \hat{w} = -(w_2 - w_1)$$

sigmoid function