Logistic Regression

used for classification

→ it outputs a probability

m instances

$x^{(i)}$

$$\begin{bmatrix} 20 \\ 95 \\ \vdots \end{bmatrix} \quad n \text{ features} \quad x_j^{(i)}$$

Money

age

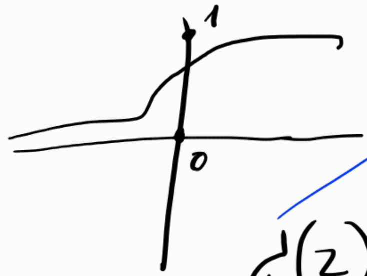
time

parameters

n features

$$\theta^T x^{(i)} = z^{(i)} \quad \text{logits}$$

Sigmoid



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

$$\underbrace{h_\theta^{(i)}}_{\text{estimated probability}}(x^{(i)}) = \sigma(z^{(i)})$$

likelihood

ground truth

$$L(\theta) = \prod_{i=1}^{m} \left[h_\theta(x^{(i)})\right]^{y^{(i)}} \left[1 - h_\theta(x^{(i)})\right]^{1-y^{(i)}}$$

1. should be negated (gradient descent) → · -1
2. Unstable numerically → ln → natural log
3. → $\frac{1}{m}$ · ......

because of log

because multiplication becomes addition

$\ln(\mathcal{L}(\theta)) =$

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \cdot \ln(h_\theta(x^{(i)})) + 1 - y^{(i)} \cdot \ln(1) - \ln(h_\theta(x^{(i)}))$$

$$J(\theta) = \frac{-1}{m} \sum_{i=1}^{m} y^{(i)} \cdot \ln(h_\theta(x^{(i)})) + 1 - y^{(i)} \cdot \ln(1) - \ln(h_\theta(x^{(i)}))$$

cross entropy loss

$j \rightarrow$ individual parameter

∂ partial derivative

$$\frac{\partial J}{\partial \ell} = -\frac{1}{m} \rightarrow \text{since } \boxed{J = \frac{-1}{m} \cdot \ell(\theta)}$$

$$\frac{\partial \ell}{\partial \theta_j} = \underbrace{\frac{\partial \ell}{\partial h}}_{y \cdot \frac{1}{h} + (1-y) \cdot \left(-\frac{1}{\ldots}\right)} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial \theta_j}$$

$$= \frac{y}{h} - \frac{1-y}{1-h} \qquad h\cdot(1-h) \qquad x_j$$

$$= \left[\frac{y}{h} - \frac{1-y}{1-h}\right]\left[h(1-h)\right] x_j^{(i)}$$

$$= (y-h)\cdot x_j^{(i)}$$

$$\frac{\partial J}{\partial \theta_j} = \frac{-1}{m}\left[(y-h)\, x_j^{(i)}\right)$$

$$\nabla_\theta J(\theta) = \frac{1}{m} X^T(h(\theta x) - y)$$

gradient