Classification & Regression

Probabilistic View of Linear Regression

Classification

Why not linear Regression?

logistic Regression

<u>METHOD</u>: Newton's METHOD

Recall Least Squares

Given $\{(x^{(i)}, y^{(i)})$ for $i = 1 \dots n\}$

in which $x^{(i)} \in \mathbb{R}^{d+1}$, $y^{(i)} \in \mathbb{R}$

Do find $\theta \in \mathbb{R}^{d+1}$ s.t. $\theta = \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)}))^2$

where $h_\theta(x) = \theta^T x$

Why?

Assume $y^{(i)} = \theta_*^T x^{(i)} + \epsilon^{(i)}$

$\hookrightarrow$ Error or noise term.

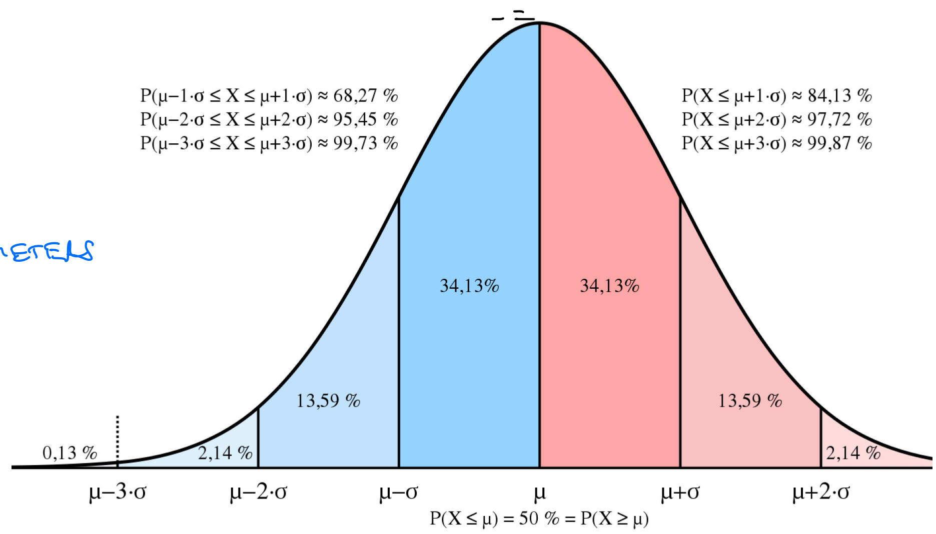
Properties

1. $\mathbb{E}[\epsilon^{(i)}] = 0$ – It's unbiased for $i \neq j$

2. The errors independent $\mathbb{E}[\epsilon^{(i)} \epsilon^{(j)}] = \mathbb{E}[\epsilon^{(i)}] \, \mathbb{E}[\epsilon^{(j)}]$ for $i \neq j$

How "Noisy" $\mathbb{E}[(\epsilon^{(i)})^2] = \sigma^2$

Gaussian or Normal Distribution (Unique of the Above)



Write $\epsilon^{(i)} \sim N(\underbrace{\cancel{\mu}}_{\mu = 0 \text{ MEAN}}, \underbrace{\sigma^2}_{\text{VARIANCE}})$

$P(z; \mu, \sigma^2) =$ PARAMETERS

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}$$

$$\exp(x) = e^x$$

$$P(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y^{(i)} - \theta \cdot x^{(i)})^2}{2\sigma^2}\right\}$$

$\epsilon^{(i)}$

θ → PARAMETER

$$y^{(i)} \mid x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)$$

Picking θ ⇒ Picks distribution

Likelihood among many distributions, "most likely"

$$\mathcal{L}(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \quad \text{(iid)}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{(y^{(i)} - \theta \cdot x^{(i)})^2}{2\sigma^2}\right\}$$

log likelihood

Depends on a data & θ.

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \left(\log \frac{1}{\sigma\sqrt{2\pi}} - \frac{(y^{(i)} - \theta \cdot x^{(i)})^2}{2\sigma^2} \right)$$

$$J(\theta) = \max_{\theta} \ell(\theta) = \min_{\theta} \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta \cdot x^{(i)})^2$$

Likelihoods

Among many distributions, Pick most likely one

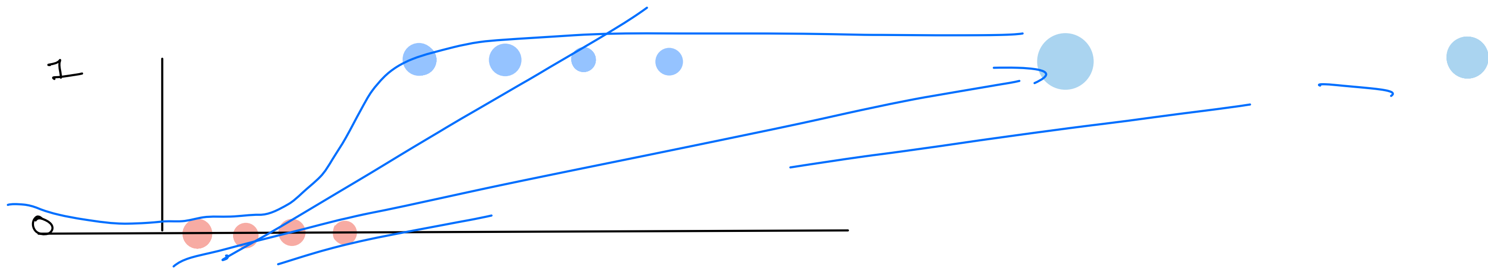
$$\mathcal{L}(\theta) =$$

Classification

Given $(x^{(i)}, y^{(i)})$ for $i = 1 \dots n$ $\quad y^{(i)} \in \{0, 1\}$

Positive class

Negative class



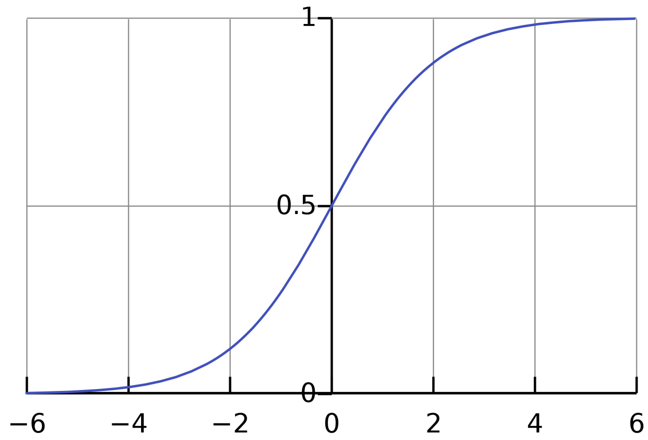
Same Recipe as Linear Regression!

$h_\theta(x) \in [0,1]$

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

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

SIGMOID



$$\boxed{P(y=1 \mid x;\theta) = h_\theta(x)}$$

$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$

likelihood

$$\mathcal{L}(\theta) = P(\vec{y} \mid \vec{x};\theta) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)};\theta)$$

$$= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}$$

log likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

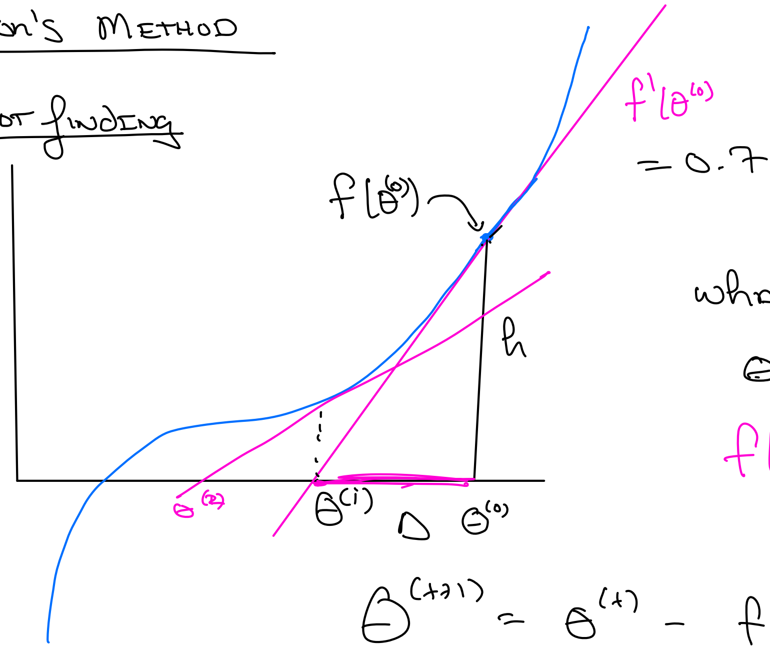
Same Recipe: $\theta^{(t+1)} = \theta^{(t)} + \alpha \frac{\partial}{\partial \theta_i} J(\theta)$

$$\frac{\partial}{\partial \theta_j} \ell(\theta) = \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)}))\, x_j^{(i)}$$

Rule is very general.

Newton's Method

Root finding



Given $f: \mathbb{R}^d \to \mathbb{R}$

Do find $f(x) = 0$

(Aside $\min \ell(\theta) \Rightarrow \ell'(\theta) = 0$)

what is Δ?

$$\theta^{(1)} = \theta^{(0)} - \Delta$$

$$f(\theta^{(0)}) = f'(\theta^{(0)}) \cdot \Delta$$

$$\Delta = f'(\theta^{(0)})^{-1} \cdot f(\theta^{(0)})$$

$$\theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$$

Error $0.1 \longrightarrow 0.01 \longrightarrow 0.0001$

Generalize & use for min. $\theta \in \mathbb{R}^{d+1}$ $\quad \ell'(\theta) = f(\theta)$

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1} \nabla_\theta \ell(\theta)$$

$\nabla_\theta \ell(\theta) \in \mathbb{R}^d$

H: Hessian $\in \mathbb{R}^{(d+1) \times (d+1)}$

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta)$$

TO find minimum,

Rough Comparison

METHOD	Per iteration	Compute	Steps to Error ϵ^{-1}
SGD	1 Data Point	$\Theta(d)$	ϵ^{-2}
Batch GD	N Data Points	$\Theta(nd)$	$\approx \epsilon^{-1}$
Newton Method	N Data points	$\Omega(nd^2)$	$\approx \log\left(\frac{1}{\epsilon}\right)$

MINIBATCH

In Classical Stats d was small

n was moderate