Classification & Regression

Probabilistic View of Linear Regression

Classification

Why not linear Regression?

logistic Regression

METHOD: Newton's METHOD

⊕ converges quickly → fast

⊖ each step is expensive

Recall Least Squares

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

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

Do Find $\theta \in \mathbb{R}^{d+1}$ s.t. $\theta = \arg\min_{\theta} \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)}))^2$

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

Why? Assume $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$ ← random variable for each i

↳ error || noise term

= stuff we can't really

↑ plug-in

1. $E[\varepsilon^{(i)}] = 0$ … it's unbiased
2. The errors independent $E[\varepsilon^{(i)}\varepsilon^{(j)}] = E[\varepsilon^{(i)}]E[\varepsilon^{(j)}]$ for $i \neq j$

How noisy $E[(\varepsilon^{(i)})^2] = \theta^2$

(unique of the above)

Gaussian or Normal Distribution

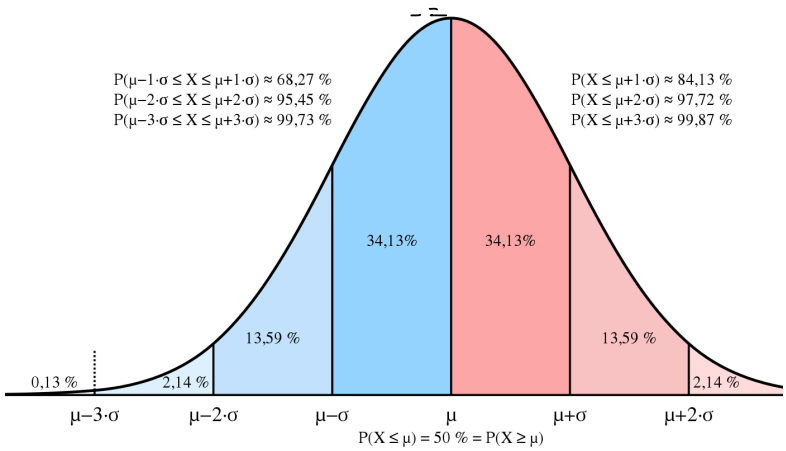
mean

Write $\epsilon^{(i)} \sim N(\mu, \sigma^2)$

plug in → variance → parameters

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

Not conditional (density)



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

parameter

histograms → eventually converges to this 정규분포

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

Picking θ ⇒ Picks distribution

Likelihood among many distributions. "most likely"

$L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; \theta)$ (iid) "independently and identically distributed"

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

$\exp(x) = e^x$

lg Likelihood

maximize $L(\theta)$ ⇔ maximize $\log L(\theta)$

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

X depend on data (θ)

depends on data (θ)

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

Least squares?

Likelihoods — Among many distributions, Pick most likely one

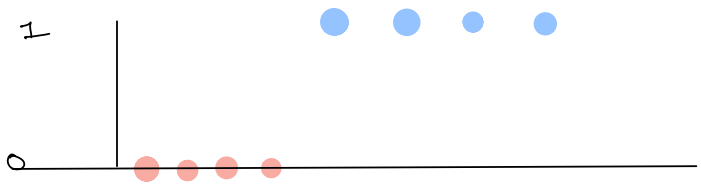
$\mathcal{L}(\theta) =$

Classification

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

$y^{(i)} \in \{0, 1\}$

- 1 → positive class
- 0 → negative class



SAME RECIPE AS linear Regression!

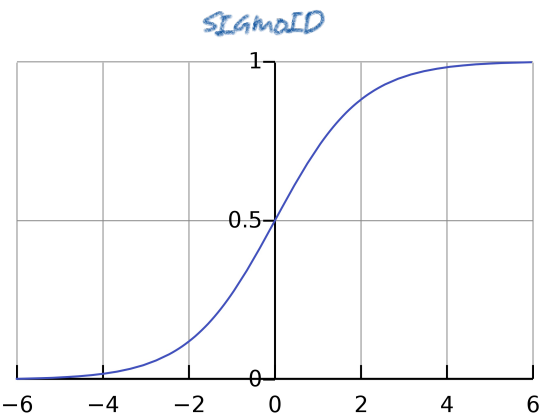
prediction

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

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

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

Homework



① Smooth (nice transition)

②

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

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

Likelihood (probability)

$L(\theta) = P(\vec{y} \mid \vec{x}; \theta)$

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

encoding

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

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

label

log likelihood

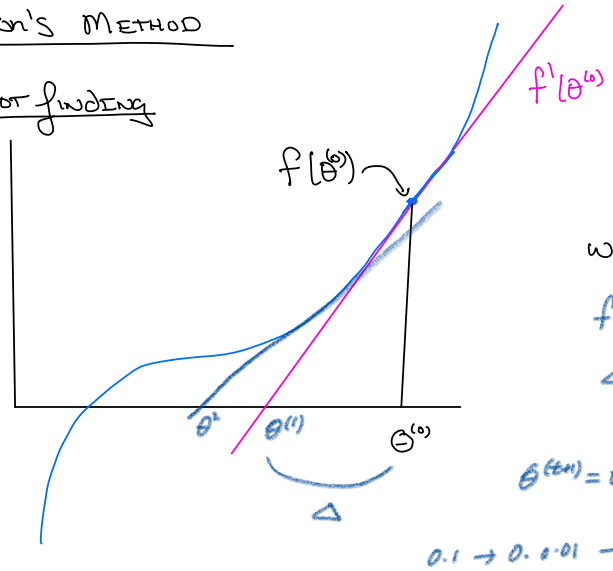
SAME RECIPE: $\theta^{(t+1)} = \theta^{(t)} + \alpha \frac{\partial}{\partial \theta_i} J(\theta)$

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

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

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

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

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

$0.1 \to 0.01 \to 0.001$

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1} \nabla_\theta \ell(\theta) \quad \in \mathbb{R}^d$$

$\theta \in \mathbb{R}^{d+1}$ $\ell'(\theta) = f(\theta)$

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

$$\text{Hessian} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \cdot \ell(\theta)$$

↑ Differentiate

To find maximum,

Stochastic Gradient Descent
확률적경사하강법
(확률적)

Rough Comparison

METHOD	Per iteration	Compute	Steps to Error ϵ^{-1}	Robustness against error
SGD	1 Data point	$O(d)$	ϵ^{-2}	robust
Batch GD	N data point	$O(nd)$	ϵ^{-1}	
Newton's method	N data points	$\Omega(nd^2)$	$\sim \log(\frac{1}{\epsilon})$ → ϵ^2 만큼빠르게 줄어듦	

↓ mini batch (parallelism)
batch는 cluster size denominator로 나눠줌