

Classification  $\neq$  Regression

Linear regression  $\rightarrow$  Probabilistic Interpretation

Classification

Why not linear regression?

Logistic Regression

Method: Newton's Method

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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  $\theta \in \mathbb{R}^{d+1}$  st.  $\theta = \operatorname{argmin}_{\theta} \sum_{i=1}^n (y^{(i)} - h_{\theta}(x^{(i)}))^2$

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

Why?

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

↳ Errors, unmodeled effects, random noise

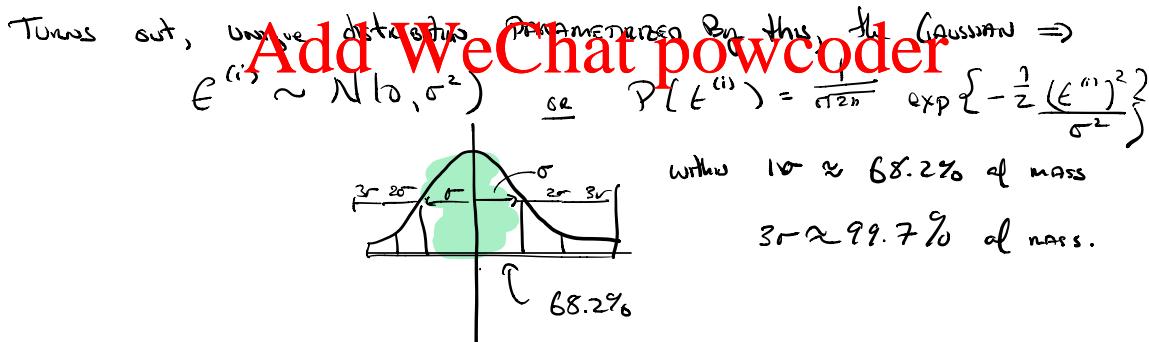
Properties of  $\epsilon^{(i)}$  we want (iid Gaussian)

1.  $\mathbb{E}[\epsilon^{(i)}] = 0$  IT'S UNBIASED

2. THE ERRORS ARE INDEPENDENT ( $\mathbb{E}[\epsilon^{(i)} \epsilon^{(j)}] = \mathbb{E}[\epsilon^{(i)}] \mathbb{E}[\epsilon^{(j)}]$  if  $i \neq j$ )

How noisy? VARIANCE  $\mathbb{E}[(\epsilon^{(i)})^2] = \sigma^2$

Assignment Project Exam Help  
<https://powcoder.com>



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

↳ Parametrized by  $\theta$

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

Picking  $\theta \Rightarrow$  Picks A DISTRIBUTION

Likelihoods Among many distributions, Pick "most likely" given all data

$$\begin{aligned} \mathcal{L}(\theta) &= p(y | x; \theta) \\ &= \prod_{i=1}^n p(y^{(i)} | x^{(i)}; \theta) \quad (\text{iid assumption}) \\ &= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(y^{(i)} - x^{(i)} \cdot \theta)^2}{2\sigma^2} \right) \end{aligned}$$

We use log likelihood (convenient)

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

Assignment Project Exam Help

<https://powcoder.com>

Thus, to find maximum likelihood, equivalently find

$$J(\theta) = \underset{\theta}{\operatorname{max}} \sum_{i=1}^n (y^{(i)} - x^{(i)} \cdot \theta)^2$$

Add WeChat powcoder □

## Classification

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

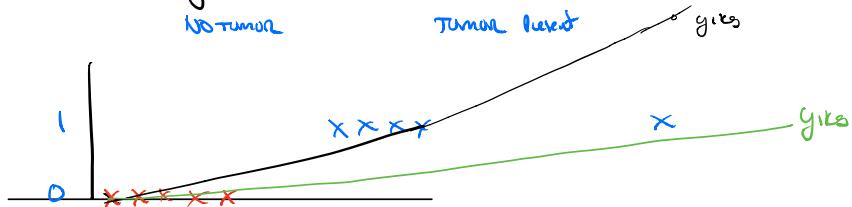
$$y^{(i)} \in \{0, 1\}$$

negative class

No tumor

Positive Class

Tumor patient



SAME Recipe

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

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

**Assignment Project Exam Help**

<https://powcoder.com>

Sigmoid or a Logistic function.

Add WeChat powcoder

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

$$P(y=0 | x; \theta) = 1 - h_{\theta}(x)$$

$$\begin{aligned} L(\theta) &= P(\vec{y} | x; \theta) = \prod_{i=1}^n P(y^{(i)} | x^{(i)}; \theta) \\ &= \prod_{i=1}^n h_{\theta}(x^{(i)})^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{(1-y^{(i)})} \end{aligned} \quad \text{"Encodes y"}$$

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

SAME Recipe maximize with gradient ascent

$$\theta_j := \theta_j + \alpha \frac{\partial \ell(\theta)}{\partial \theta_j} \quad (\text{Gradient Ascent})$$

$$+ + \theta_0$$

of last week  $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$  (least squares)

fun observation

$$\frac{\partial J(\theta)}{\partial \theta_j} = \{y^{(i)} - h_{\theta}(x^{(i)})\} x_j^{(i)} \quad \text{so ...}$$

$$\Theta := \Theta - (h_{\theta}(x) - y) x$$

We'll see later this rule is very general.

## Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## Newton's Method

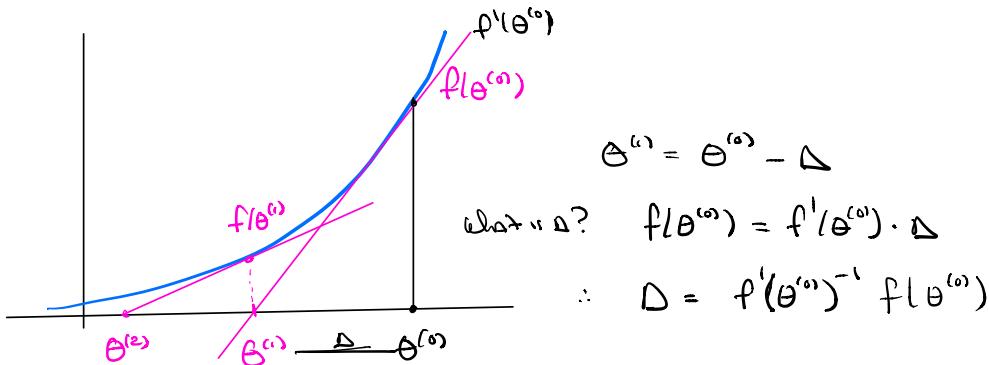
Given  $f: \mathbb{R}^d \rightarrow \mathbb{R}$

Do find  $x$  s.t.  $f(x) = 0$ .

## ASIDE

$\max_{\theta} l(\theta)$  want  $l'(\theta) = 0$

↑  
derivative



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

## Assignment Project Exam Help

Converges fast! Quadratic  $0.1 \rightarrow 0.01 \rightarrow 0.0001$  (digit double!)

Generalizing to vector  $\theta \in \mathbb{R}^{d+1}$  and  $l'(\theta) = f(\theta)$  (i.e. minimization)

Add WeChat  powcoder

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

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

$\hookrightarrow \in \mathbb{R}^{d+1}$

NB: NO STEPSIZE!

## Rough Comparison

SGD

Per Iteration

1 DATA point

Compute

$O(d)$

STEPS TO ERROR E

$E^{-2}$

Batch SGD

n DATA points

$O(nd)$

$\approx E^{-1}$

Newton

n DATA points

$O(n^2)$

$\log(\frac{1}{E})$

IN Classical stats  $d$  is small 100 or so

And Exact Answer matters  $\Rightarrow$  Newton (LBFGS)

modern ML

$d$  is Huge 1GB  $d^2 \Rightarrow n$

$\Rightarrow$  SGD of ten WORKHOUSE (exact solution less 11)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder