Classification & Regression

Linear regression → Probabilistic Interpretation

Classification

Why not linear regression?

logistic Regression

METHOD: Newton's Method

Recall Least Squares

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

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

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

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

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

$\hookrightarrow$ Error, unmodeled effects, random noise

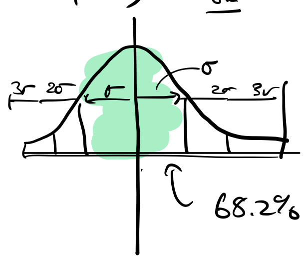
Properties of $\epsilon^{(i)}$ we want (iid Gaussian)

1. $\mathbb{E}[\epsilon^{(i)}] = 0$, it's unbiased
2. the errors are independent $(\mathbb{E}[\epsilon^{(i)}\epsilon^{(j)}] = \mathbb{E}[\epsilon^{(i)}]\mathbb{E}[\epsilon^{(j)}] \quad i \neq j)$

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

Turns out, unique distribution parametrized by this, the Gaussian $\Rightarrow$

$\epsilon^{(i)} \sim N(0, \sigma^2)$ or $P(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\epsilon^{(i)})^2}{\sigma^2}\right\}$



within $1\sigma \approx 68.2\%$ of mass

$3\sigma \approx 99.7\%$ of mass.

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

$\hookrightarrow$ Parametrized by θ

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

Picking $\theta \Rightarrow$ Picks a distribution

Likelihoods

among many distributions, pick "most likely" given all data

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

We use log likelihood (convenient)

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

Does depend on θ

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

↳ Doesn't Depend on θ

Thus, to find maximum likelihood, equivalently find

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

□

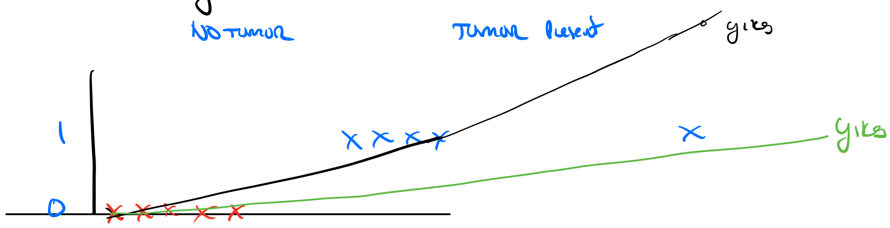
Classification

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

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

negative class — NO TUMOR

Positive Class — TUMOR Present

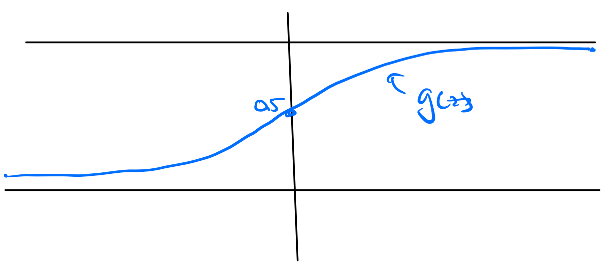


Same Recipe

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

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

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



Sigmoid or a logistic function.

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

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

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

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

"Encodes if"

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

Same Recipe maximize with gradient Ascent

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

$\partial \theta_j$

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

Fun observation

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)} \quad \text{so} \ldots$$

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

We'll see later this rule is very general.

Newton's Method

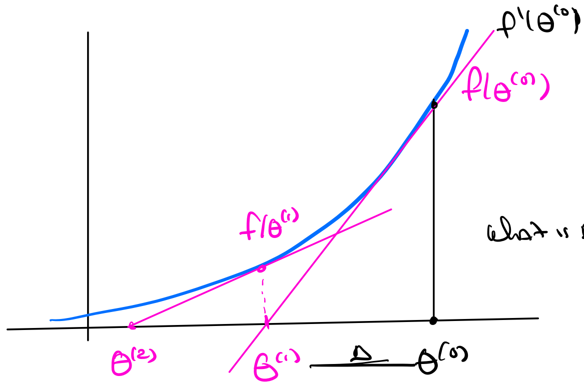
Given $f : \mathbb{R}^d \to \mathbb{R}$

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

Aside

$\max_\theta \ell(\theta)$ want $\ell'(\theta) = 0$

↑ derivative



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

what is Δ? $f(\theta^{(0)}) = f'(\theta^{(0)}) \cdot \Delta$

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

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

Converges fast! Quadratic $0.1 \longrightarrow 0.01 \longrightarrow 0.0001$ (digits double!)

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

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

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

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

NB: No stepsize!

Rough Comparison	Per iteration	Compute	Steps to Error ϵ
SGD	1 data point	$O(d)$	ϵ^{-2}
Batch SGD	n data points	$O(nd)$	$\approx \epsilon^{-1}$
Newton	n data points	$\Omega(nd^2)$	$\log(\frac{1}{\epsilon})$

In classical stats d is small 100 or so

And exact answer matters $\Rightarrow$ Newton (LBFGS)

modern ML d is HUGE 175B $d^2 \Rightarrow$ ☹

$\Longrightarrow$ SGD OFTEN WORKHORSE (Exact solution less so]