Lecture -3

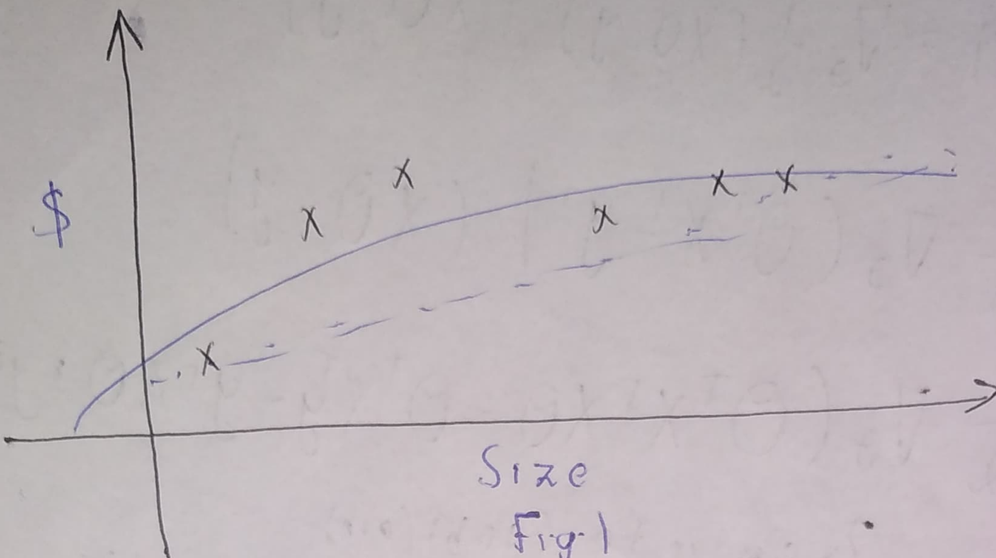


Fig 1

- You could define new feature to increase the complexity of your hypothesis
  - $\theta_0 + \theta_1 x_1$
  - $\theta_0 + \theta_1 x_1 + \theta_2 x_2^2$
  - $\theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_2}$

Locally Wieghtage Linear Regression:-

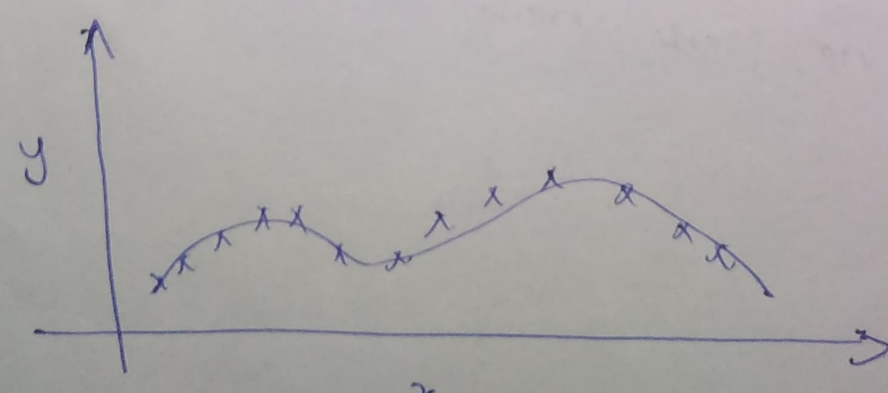


Fig 2

(*) Difference between paramatric and non-parametric method

→ In $\boxed{\text{parmateric}}$ learning algorithm → You fit fixed set Θ_i to data

→ In non-paramateric:-

• Amount of data/parameters you need to keep grow with size of data.

→ Locally Wieghted Regression is non-parametric.

(*)

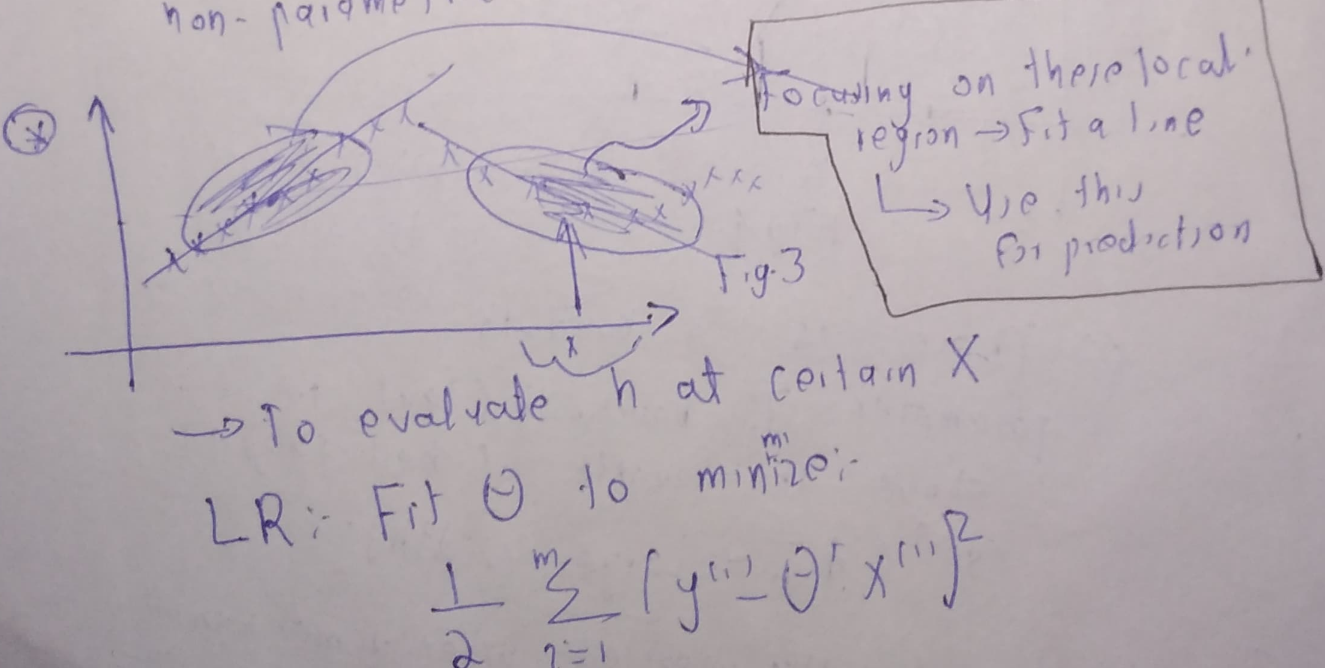


Fig.3

→ To evaluate h at certain X

LR:- Fit Θ to minimize:-

$$\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\theta^T x^{(i)}\right)^2$$

Return $\theta^T X$.

- For locally wieghted regression:-

Fit Θ to minimize

$$\sum_{i=1}^{m} \boxed{w^{(i)}} \left(y^{(i)} - \Theta^T x^{(i)}\right)^2 \quad \text{Eq.1}$$

where $w^{(i)}$ is a wieght function

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

Eq.2

- If $|x^{(i)} - x|$ is small, $w^{(i)} \approx 1$ else it will be close to zero

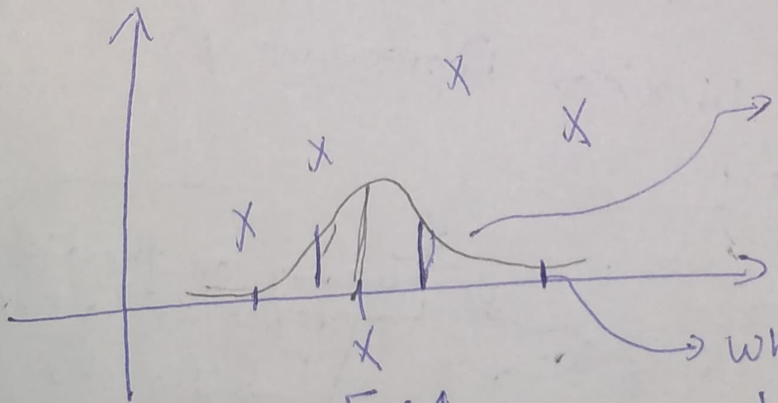
- 

Fig.4

- There is a bandwith parameter τ = bandwith (Tau) to choose the width of bell-shape curve ($\uparrow \tau \rightarrow$ Larger width).

- Use locally wieghted when dimension of your dataset is small and you have a lot of data.

Probalitistic Intrepretation of Linear Regression:-

(~~Why Least~~

Why Least Squares?

→ Assume ① $y^{(i)} = \theta^T \cdot x^{(i)} + \varepsilon^{(i)}$

These independently identically distributed (IID).

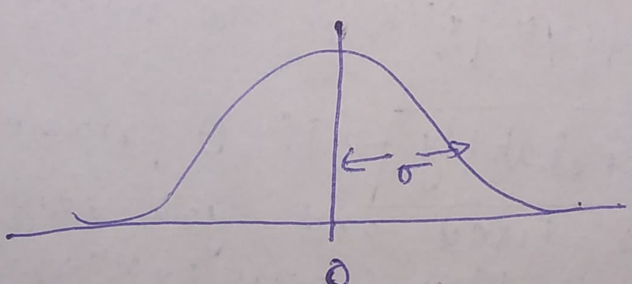
error term : unmodelled effects, random noise.

② $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$.

~~P(ε~~

→ $P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right)$

↳ PDE for $\varepsilon^{(i)}$



→ This implies:-

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

16

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

"parmaterized as by."

$\mathcal{L}(\theta) = P(y | x; \theta)$.

likelihood

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

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

- What is the difference between probability & likelihood?

(1) $\rightarrow$ $\boxed{\text{Likelihood}}$ $\rightarrow$ If you view $P(y^{(i)} | x^{(i)}; \theta)$ as a function of θ then $P(y^{(i)} | x^{(i)}; \theta)$ is likelihood. (Data is fixed)

(2) $\rightarrow$ $\boxed{\text{Probability}}$ $\rightarrow$ If θ parameter is fixed and you vary the data then $P(y^{(i)} | x^{(i)}; \theta)$ is probability.

→ Log likelihood:-

$$\ell(\theta) = \log L(\theta)$$

$$= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp(\cdots)$$

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

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

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

(*) See this equation → the black part shows the MSE Formula so maximize log-likelihood is equivilent to minimizing MSE i.e $\frac{1}{2}\sum_{i=1}^{m}(y^{(i)}-\theta^T\cdot x^{(i)})^2$
↳ $J(\theta)$ (Cost function).

→ Maximum Likelihood Estimate:-

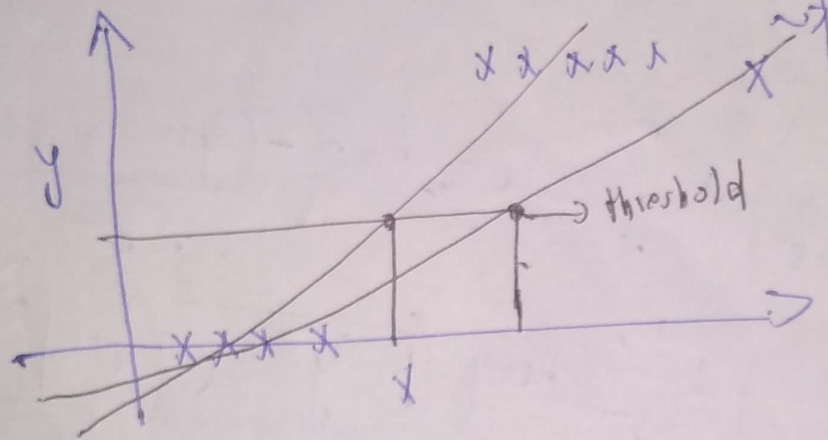
→ Choose θ to ~~maximize~~ $L(\theta)$ → maximize likelihood.

↳ Easier to ~~m~~ maximize log-likelihood ($\ell(\theta)$)

So ~~in Linear Regression choose~~
~~θ to minimize~~

Classification:-

- $y \in \{0,1\}$ (binary classification).



Linear Regression won't work because it is suspectible to extreme value ~~any~~ ge and generally won't fit given the pattern of dat1

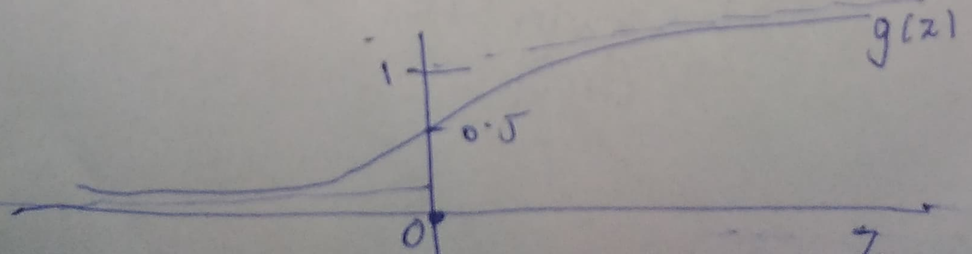
Logistic Regression:-

→ Want $h_\theta(x) \in [0,1]$

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

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

↳ Sigmoid or logistic function.



- Why $g(z) = \frac{1}{1+e^{-z}}$ because it then form generalized linear model (Explanation in future Lectures).

- Let's make some $\boxed{\text{assumption}}$:

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

  ② $\boxed{P(y=1 \mid x;\theta) = 1 - h_\theta(x)}$

  ③ $\boxed{y \in \{0,1\}}$

Putting / Compressed in one equation

- $P(y \mid x;\theta) = \cancel{y}\,(h_\theta(x))^y \cdot (1-h_\theta(x))^{(1-y)}$.

$\boxed{\begin{array}{l} \text{If } y=1:- P(y \mid x;\theta) = (h_\theta(x))^y \\ \text{if } y=0:- P(y \mid x;\theta) = 1-h_\theta(x) \end{array}}$

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

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

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

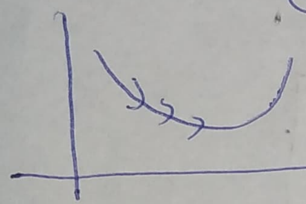
$$\overset{\log}{=} \sum \prod_{i=1}^{m} (h_\theta(x))^{y^{(i)}} \cdot (1-h_\theta(x))^{1-y^{(i)}}).$$

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

$$\boxed{\text{Choose } \theta \text{ to maximize } \ell(\theta)}$$

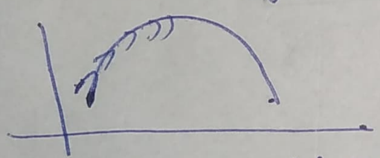
* Batch gradient descent to choose θ:

$$\boxed{\theta_j := \theta_j + \alpha \frac{\partial}{\partial \theta_j} \ell(\theta)}$$



- sign (descent)

↳ Plus sign because we are try to maximize log-likelihood



+ sign (Asent)

$$\theta_j := \theta_j + \alpha \underbrace{\left(\sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right)\right) x_j^{(i)}}_{\frac{\partial \ell(\theta)}{\partial \theta_j}}$$

$\ell(\theta)$ is always concave so there is global optima.

→ Same update as linear regression but $h_\theta(x^{(i)})$ is now different.

Newton Method:-

- Gradient ascent takes a lot of baby steps to converge.
- N.M takes bigger step and need fewer iteration to converge.

One-dimensional Justification for NM :-

- Have f
- Want to find θ s.t $f(\theta) = 0$.
- Want to maximize $\ell(\theta)$
  [Want $\ell'(\theta) = 0$
  i.e want $\ell'(\theta) = 0$]
  ↳ we will $f(\theta) = \ell'(\theta)$ in L.R

But
- We will work with this example to build our understanding.

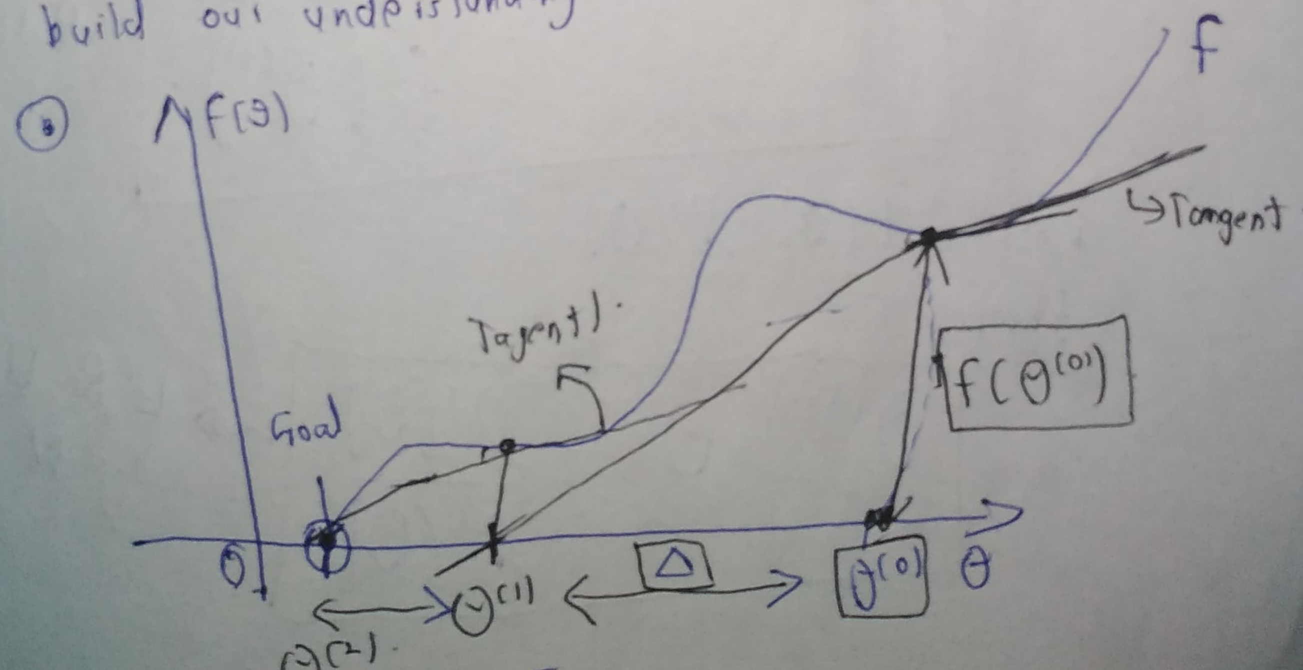


Fig.

Maths for NM

* From above fig:-

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

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

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

* $$\theta^{(1)} = \theta^{(0)} - \frac{f(\theta^{(0)})}{f'(\theta^{(0)})}.$$

Generally

$$\boxed{\theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}.}$$ → Update Rule for NM.

Let $f(\theta) = \ell'(\theta)$

then:-

$$\boxed{\theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta)}{\ell''(\theta)}.}$$ → L.R Update by N.M.

(*) Newton Methods enjoys the property of $\boxed{\text{Quadratic Convergence:}}$

→ After one iteration the error of margin from optimal value i.e $f(\theta)=0$ decrease quadratically (double):-

(e) 0.01 error → 0.0001 error → 0.00000001 error

→ Reason why N.M requires fewer iteration as compare to L.R.

(*) $\boxed{\text{NM update Rule}}$ when Θ is a Vector:

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

Matrix of $\mathbb{R}^{n+1 \times n+1}$

Vector $\mathbb{R}^{n+1}$ of derivatives

In high-dimensions Newton Method is however computationally expensive because each iteration requires inverting a large matrix

- where H is the Hessian matrix
- Hessian matrix is matrix of partial derivatives:

$$\boxed{H_{ij} = \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}}$$

Rule of Thumb

→ Less parameters → N.M

→ More parameters → G.D