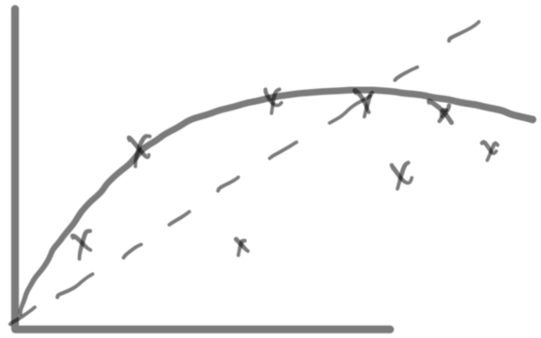
Stanford CS229 lecture-2

Locally Weighted regression:

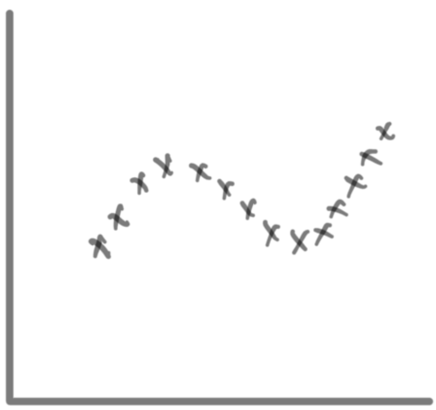


- - - - $\theta_0 + \theta_1 x + \theta_2 x$

—— $\theta_0 + \theta_1 x + \theta_2 \sqrt{x}$

As we can see, this feature fits the data better

But



→ ? How do we fit curve onto this?

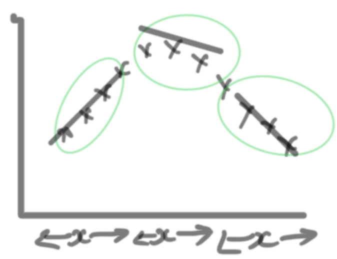
Linear Regression

Goal: Fit θ to minimize

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

return $\theta^T x$

Locally Weighted regression



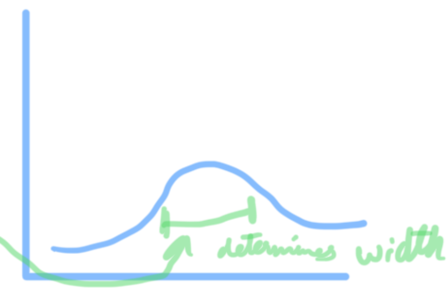
Goal: Find straight line or curve to fit a local set of points i.e. fit θ to minimize

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

where, $w^{(i)} = e^{\left(\frac{-(x^{(i)}-x)^2}{2}\right)}$

$\lvert x^{(i)}-x \rvert$ is small then $w^{(i)} \approx 1$ else $w^{(i)} \approx 0$	$x^{i} \rightarrow$ set of all neighbour points; $x \Rightarrow$ the point at which we are deducing LWR

$e^{\frac{-(-x^{i}-x)^2}{2}}$



→ if x^{i} is close to x greater weight

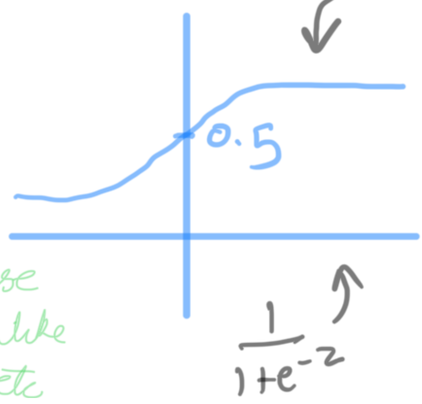
Logistic Regression

↳ Classification

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

Intuition

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



- We specifically choose sigmoid
  - → give (0,1) output
  - → non linear
  - → differentiable
    - ↳ i.e we can use optimization algos like gradient descent etc

Probabilistic formula

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

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

$y \rightarrow$ output
$x \rightarrow$ input
$\theta \rightarrow$ parameter

Combining these 2 formulas

$$P(y \mid x;\theta) = h_\theta(x)^y \left(1 - h_\theta(x)\right)^{1-y}$$

if $y=1$ $(1-h(x))^{1-1} \rightarrow 0$
if $y=0$ $h(x)^{0} \rightarrow 0$

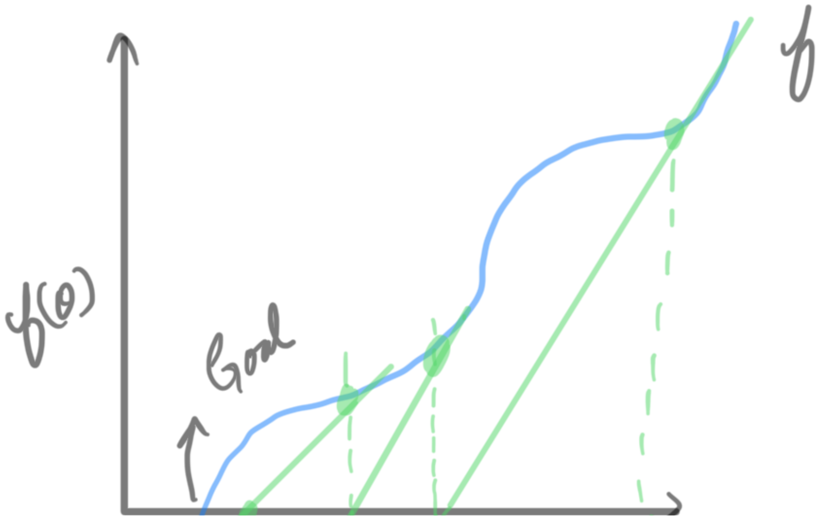
Newton's Method

↳ Alternative to Gradient descent
• For less parameters

Goal :-

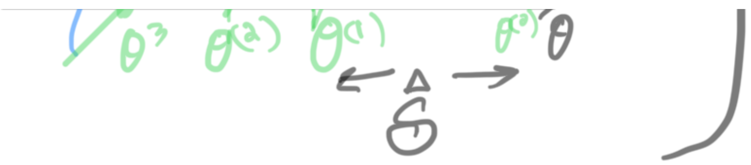
maximize some function $\ell(\theta)$
↳ means $\ell'(\theta) = 0$

let us say $\ell'(\theta) = f(\theta)$
$\rightarrow$ finding θ so that $f(\theta) = 0$
i.e $\ell'(\theta) = 0$



Algorithm :-

• Take a random point
• go θ, where the tangent (dy/dx) touches it.
• repeat



Math behind this

$$\theta^{(2)} := \theta^{(0)} - \Delta$$ (We have to solve for Δ)

$$f'(\theta^0) = \frac{f(\theta^0)}{\Delta} \quad \left[\frac{dy}{dx} = \frac{\text{height}}{\text{width}}\right]$$

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

i.e

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

Since $f(\theta) = l''(\theta)$

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

"Quadratic Convergence."

↓

Informally means,

0.01 error → 0.0001 error → 0.00000001 error

↳ Newton's method thus for each iteration converges faster.