Outline

- Linear Regression (recap)
- Locally Weighted Regression
- Probabilistic interpretation
- Logistic Regression
- Newton's Method

Recap:

$(x^{(i)}, y^{(i)})$ i^{th} example

$x^{(i)}$: features for i^{th} training example

$x^{(i)} \in \mathbb{R}^{d+1}$ $y^{(i)} \in \mathbb{R}$ $x_0 = 1$

n: # examples, d = # features

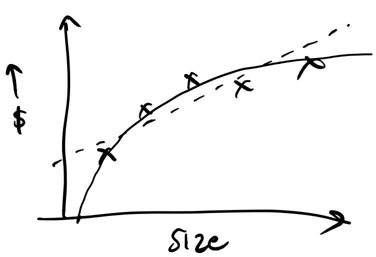
$h_\theta(x^{(i)})$: prediction on i^{th} training example

$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = \theta^T x$$

x : feature vector new example

$h_\theta(x)$

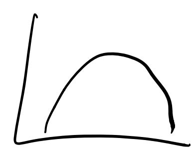
$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

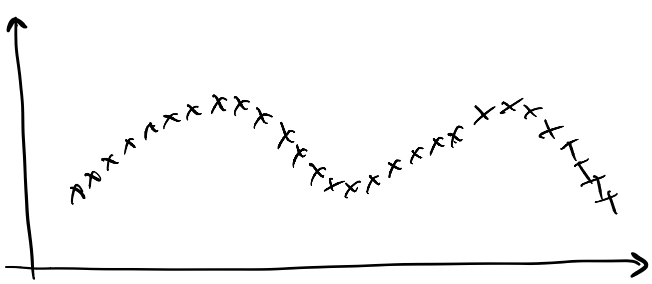


$\theta_0 + \theta_1 x_1$

$\theta_0 + \theta_1 x + \theta_2 x^2$

$\theta_0 + \theta_1 \underset{\uparrow \atop x_1}{x} + \theta_2 \underset{\uparrow \atop x_2}{\sqrt{x}} + \theta_3 \underset{\uparrow \atop x_3}{\log(x)}$





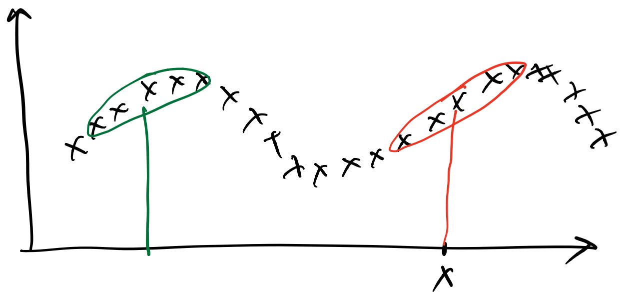
Locally Weighted Regression

"Parametric" learning algorithm

Fit fixed set of parameters (θ_i) to data

"Nonparametric" learning algorithm

\# parameters grows linearly with size of data



x: feature vector for new example

To evaluate h at certain x

LR: Fit θ to minimize

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

Return $\theta^T x$

Locally Weighted Regression

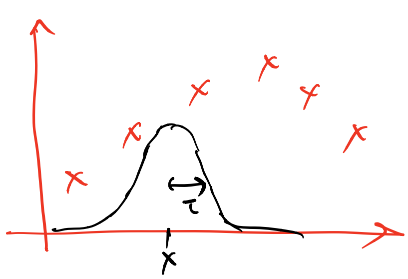
Fit θ to minimize

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

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

If $|x^{(i)} - x|$ small $\quad w^{(i)} \approx 1$

$|x^{(i)} - x|$ large $\quad w^{(i)} \approx 0$



τ: bandwidth

Probabilistic interpretation

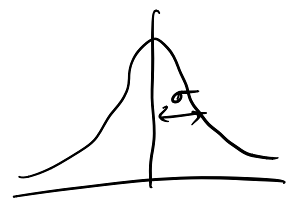
Why Least Squares?

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

"error": unmodeled effects, random noise

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

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



Assumption: $\varepsilon^{(i)}$ are IID

This implies that

random variable

mean

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

"parametrized by"

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

$P(y^{(i)} \mid x^{(i)}, \theta)$

$$\mathcal{L}(\theta) = p(\vec{y} \mid X ; \theta)$$
$$= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)} ; \theta)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right)$$

Log likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta)$$
$$= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp(\quad)$$

$$= \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi}\,\sigma} + \log \exp(\quad) \right]$$

$$= n \log \frac{1}{\sqrt{2\pi}\,\sigma} + \sum_{i=1}^{n} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}$$

MLE : Maximum Likelihood Estimation

Choose θ to maximize $\mathcal{L}(\theta)$

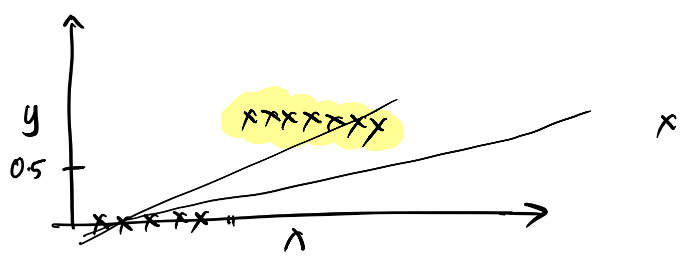
Minimize $\frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 = J(\theta)$

Classification :

① Make assumption $P(y|X;\theta)$

② Compute θ by MLE

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



Logistic Regression

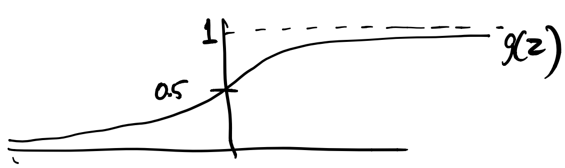
Want $h_\theta(x) \in [0, 1]$

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

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

Linear Regression

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



"sigmoid" fn or "logistic" fn

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

$y=1$: tumor malignant; x: size of tumor

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

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

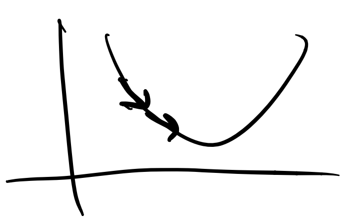
$$\begin{aligned}
\mathcal{L}(\theta) &= p(\vec{y} \mid X;\theta) \\
&= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta) \\
&= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}}
\end{aligned}$$

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

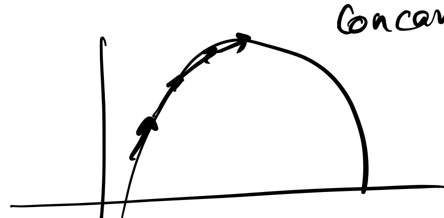
(Batch) Gradient Descent

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

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



Concave



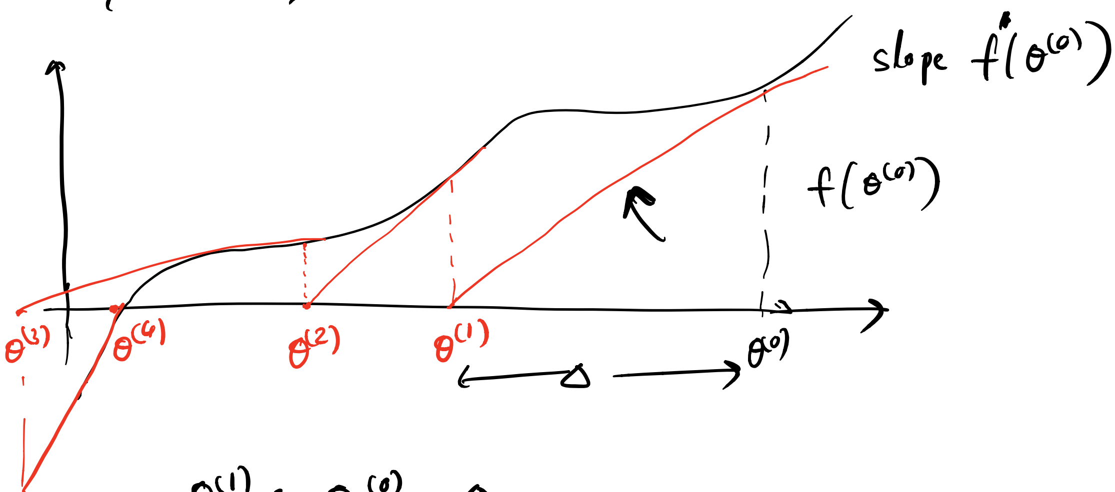
$$\theta_j := \theta_j + \alpha \underbrace{\sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)}))\, x_j^{(i)}}_{\frac{\partial}{\partial\theta_j} \ell(\theta)}$$

Newton's Method

Have f
Find θ s.t. $f(\theta)=0$

[Want: maximize $\ell(\theta)$
i.e. want $\ell'(\theta)=0$
$\uparrow$
derivative]

max/min $\iff$ derivative $=0$



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

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

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

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

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

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

Quadratic convergence

$0.1 \longrightarrow 0.01 \longrightarrow 0.0001$

θ vector

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

$$\theta^{(t+1)} := \theta^{(t)} + H^{-1} \underbrace{\nabla_\theta \ell}_{\text{vector } \mathbb{R}^{d+1}}$$

Hessian H: $\quad H_{ij} = \dfrac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}$