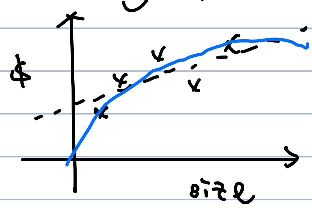
Outline

- Linear Regression (recap)
- Locally weighted Regression
- Probailistic interpretation
- Logistic Regression
- Newton's Method

Recap

- $(x^{(i)}, y^{(i)})$ i-th example
- $x^{(i)} \in \mathbb{R}^{d+1}$, $y^{(i)} \in \mathbb{R}$. $x_0 = 1$
- n: # of examples, d = # of features.
- $h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = \theta^T x$
- $J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2$

Housing prize example.



$\theta_0 + \theta_1 x$

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

$\theta_0 + \theta_1 x + \theta_2 \sqrt{x} + \theta_3 \log x$

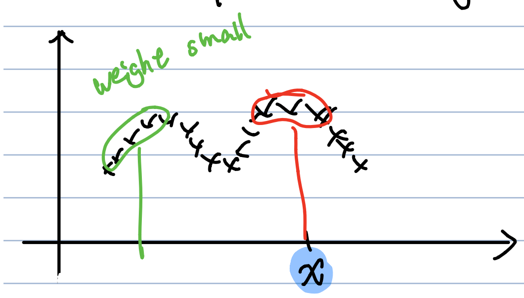
Locally Weighted Regression

"Parametric" learning algorithm.

Fit fixed set of parameters (θ_i) to data

"Nonparametric" learning algorithm

# of parameters grows linearly with size of data



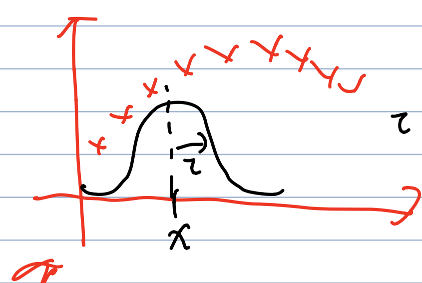
To evaluate h at certain x.

w: Weight function

LR: Fit θ to minimize

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

Return $\theta^T x$



τ: Bandwidth

Locally Weighted Regression

Fit θ to minimiz

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

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

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

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

Fit the θ everytime we predict a new x.

Probabilistic interpretation

why least squares?

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

"error": unmodeled effects, random noise

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

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

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

This implies that

random variable

mean

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

"parametrized by"

Note: $P(y^{(i)} | x^{(i)}, \theta)$ doesn't make sense. Since θ is not a random var.

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

$$\mathcal{L}(\theta) = P(\vec{y} | X; \theta)$$

likelihood of θ

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

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

Note:

Random variables $\rightarrow$ probability

Parameters (θ) $\rightarrow$ likelihood

$\mathcal{L}(\theta)$: Probability of the outcome y given the data x and para θ

Log likelihood

$$l(\theta) = \log \mathcal{L}(\theta)$$

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

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

constant

what we want to max.

Maximum Likelihood Estimation (MLE)

choose θ to maximize $L(\theta)$

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

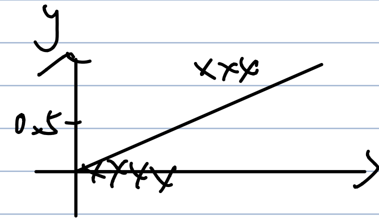
Classification

Use Linear regression.

① Make assumption $P(y|x;\theta)$

② Compute θ by MLE

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



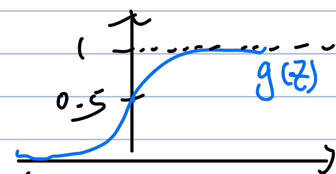
LR not ideal!

Logistic Regression

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

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

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



"Sigmoid" fn or "logistic" fn

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

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

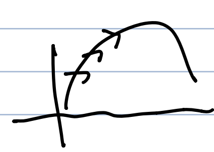
$L(\theta) = P(\vec{y}|x;\theta)$

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

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

(Batch) Gradient Descent

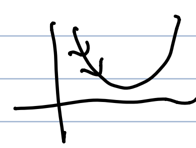
$\theta_j := \theta_j + \alpha \frac{\partial}{\partial\theta_j} l(\theta)$ Gradient Ascend



concave

(last week)

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial\theta_j} J(\theta)$ Gradient Descend



$\Rightarrow \theta_j := \theta_j + \alpha \underbrace{\sum_{i=1}^{n}\left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}}$

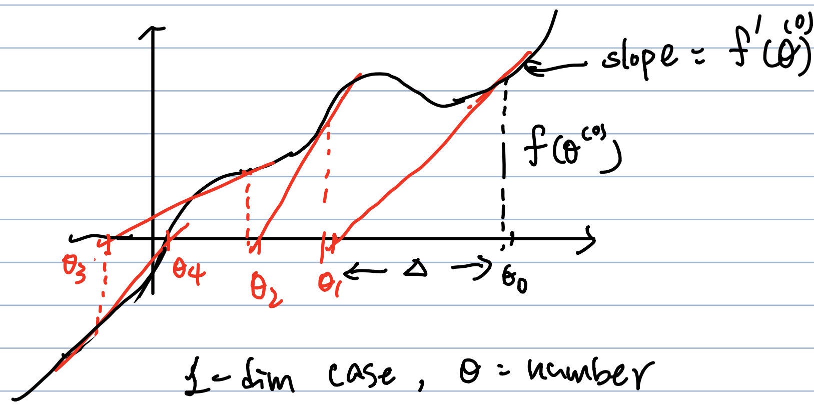
$\frac{\partial}{\partial \theta_j} \ell(\theta)$ (gradient)

Newtons Method

Have func θ s.t. $f(\theta) = 0$.

[want maximize $\ell(\theta)$
↳ want $\ell'(\theta) = 0$
derivative $= 0$]

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



1-dim case, θ = number

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

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

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

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

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

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

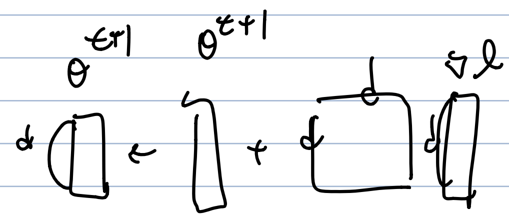
Quadratic convergence

0.1 → 0.01 → 0.0001

θ vector

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

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



Hessian H:

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

Inverse H^{-1} is expensive.
But we could converge fast!

Note. Good if dimension is low. Could compute H^{-1} easy.
Converge fast.