Lecture 8: 2/13/17

Optimization for Linear Regression

#1) Closed Form Solution:

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_M^{(1)} \\ \vdots & \vdots & & \\ 1 & x_1^{(N)} & \cdots & x_M^{(N)} \end{bmatrix} \leftarrow \vec{x}^{(N)}$$

x_0 column

"Design Matrix"

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(y^{(i)} - (\vec{\theta}^T x^{(i)}) \right)^2$$

← "Ordinary Least Squares"

$$= \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \underbrace{\frac{1}{2} (X\theta - \vec{y})^T (X\theta - \vec{y})}_{J(\theta)}$$

$$= (X^T X)^{-1} (X^T \vec{y})$$

the "Normal Equations"

$$\nabla J(\theta) = X^T X \theta - X^T \vec{y} = 0$$

Computational Complexity of Closed Form

Background: Matrix Multiplication

Matrices A and B

If A is $q \times r$ and B is $r \times s$: AB takes $O(qrs)$

If A and B are $q \times q$: AB takes $O(q^{2.373})$

If A is $q \times q$: A^{-1} takes $O(q^{2.373})$

OLS

$$(X^T X)^{-1} (X^T y)$$

X^T: $M \times N$, X: $N \times M$, $(X^T X)$: $M \times M$

X^T: $M \times N$, y: $N \times 1$, $(X^T y)$: $M \times 1$

$X^T X$	$O(M^2 N)$
$(\ )^{-1}$	$O(M^{2.373})$
$X^T y$	$O(MN)$
$(\ )(\ )$	$O(M^2)$
total	$O(M^2 N + M^{2.373})$

$\Rightarrow$ Linear in # examples, N

Polynomial in # features, M

Stability

Can we invert X^TX?

Case #1: When $N \gg M$; Yes so long as X has full column rank

Case #2: When $N \ll M$: X does not have full rank
$\Rightarrow$ problem is non-identifiable

#2) SGD for Linear Regression

Called "Least Mean Squares (LMS)" algorithm

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} J_i(\theta) \quad \text{where} \quad J_i(\theta) = \frac{1}{2}\left(y^{(i)} - \vec{\theta}^T\vec{x}^{(i)}\right)^2$$

(doesn't affect argmin)

$$\nabla J_i(\theta) = \begin{bmatrix} \frac{dJ_i(\theta)}{d\theta_1} \\ \vdots \\ \frac{dJ_i(\theta)}{d\theta_M} \end{bmatrix}$$

$$\frac{dJ_i(\theta)}{d\theta_j} = \left(y^{(i)} - \theta^T x^{(i)}\right)\frac{d}{d\theta_j}\left(y^{(i)} - \sum_{m=1}^{M}\theta_m x_m^{(i)}\right)$$

$$= -\left(y^{(i)} - \theta^T x^{(i)}\right)x_j^{(i)}$$

$$\star = -\left(y^{(i)} - \vec{\theta}^T\vec{x}^{(i)}\right)\vec{x}^{(i)}$$

(scalar) (vector)

1) Choose initial $\vec{\theta}$ (zeros, random)
2) For $t = 1 \ldots T$:
   a) Sample $i \sim \text{Uniform}(\{1, \ldots, N\})$
   b) compute $\nabla J_i(\theta)$
   c) Choose learning rate $\gamma^{(t)}$
   d) Update $\vec{\theta} = \vec{\theta} - \gamma^{(t)}\nabla J_i(\theta)$

for m in $\{1, \ldots, M\}$:
$\theta_m = \theta_m + \gamma(\quad)$

for first code it's better to use the vector version $\star$

#3 Gradient Descent for Lin. Reg.

convex

- Straightforward application of GD to solve: $\hat{\theta} = \arg\min_{\theta} J(\theta)$
- How to compute $\nabla J(\theta)$?

$$\Rightarrow \nabla J(\theta) = \sum_{i=1}^{N} \nabla J_i(\theta) = \sum_{i=1}^{N} -(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)})\, \vec{x}^{(i)}$$

Prob. Interp. of Linear Regression

Def: Generative probabilistic model is $p(x, y|\theta)$

Def: Discriminative probabilistic model is $p(y|x, \theta)$

Generative vs. Discriminative Views

	Gen	Disc.
① Assume unk. data dist.	$(x^{(i)}, y^{(i)}) \sim p^*(x, y|\theta^*)$	$(x^{(i)}, y^{(i)}) \sim p^*(x, y|\theta^*)$
② Choose model family	$p(x, y|\theta)$	$p(y|x, \theta)$
③ Learn	Maximize Likelihood ★ $\hat{\theta} = \arg\max_{\theta} \log \prod_{i=1}^{N} p(x^{(i)}, y^{(i)}|\theta)$	Maximize Cond. Likelihood. ★ $\hat{\theta} = \arg\max_{\theta} \log \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}, \theta)$
④ Predict.	★ $\hat{y} = h(x) = \arg\max_{y} p(y|x, \theta)$ $= \arg\max_{y} p(x, y|\theta)$	★ $\hat{y} = h(x) = \arg\max_{y} p(y|x, \theta)$
	★ Replace $p(x,y)$ and $p(y|x)$ with probability density functions if y continuous	★ Takeaway: $p(y|x)$ is used for predict, so just learn it directly.

Recall: Gaussian Dist.

Let X be a Gaussian r.v.: $X \sim \text{Gaussian}(\mu, \sigma^2)$

$\mu \in \mathbb{R}$ $\quad \sigma^2 > 0$

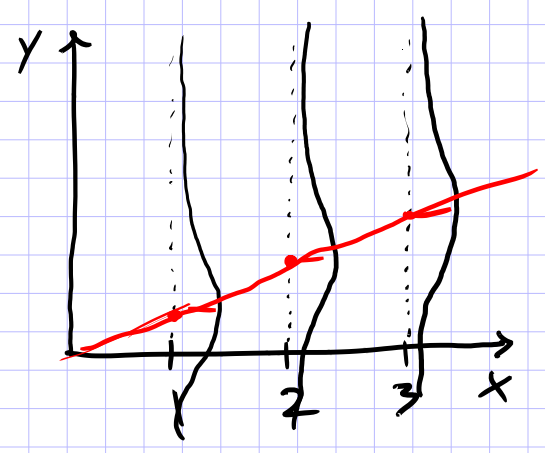
Def: probability density fn.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Case #1: 1D Lin. Reg.

Assume: two r.v.s X, Y

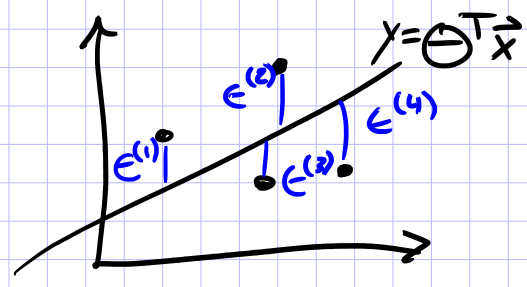
Goal: learn a function $f: X \rightarrow P(Y \mid X)$



Case #2: Multiple Linear Regression

Assume: r.vs $X_1^{(i)}, \ldots, X_M^{(i)}, Y^{(i)}$

data $D = \{\vec{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$



Story: (generate $y^{(i)}$ given $x^{(i)}$)

$$\epsilon^{(i)} \sim \text{Gaussian}(\mu = 0, \sigma^2)$$ ← 1D Gaussian

$$y^{(i)} = \vec{\theta}^T \vec{x}^{(i)} + \epsilon^{(i)}$$ ← Is $Y^{(i)}$ still a r.v.?

Equivalent Story:

$$y^{(i)} \sim \text{Gaussian}(\mu = \theta^T x^{(i)}, \sigma^2)$$

Q: What is the pdf of $y^{(i)}$?

$$g(y^{(i)} \mid x^{(i)}, \vec{\theta}, \sigma^2) = f_{\text{Gaussian}}(y^{(i)} \mid \mu = \vec{\theta}^T \vec{x}^{(i)}, \sigma^2)$$

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

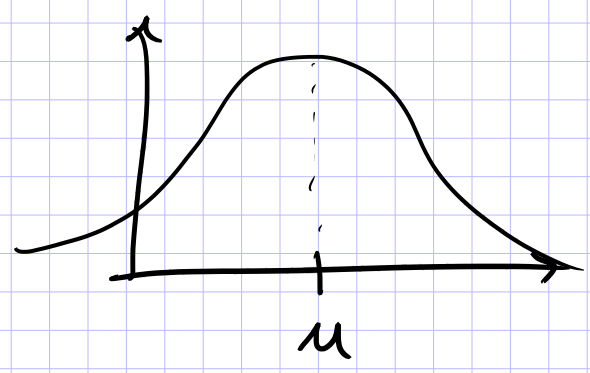
Predictions

use pdf of y since its continuous

$$\hat{y} = h(\vec{x}) = \underset{y \in \mathbb{R}}{\operatorname{argmax}}\ g(y \mid \vec{x}, \vec{\theta}, \sigma^2)$$

The mode of a Gaussian is its mean.

(most probable value)



$$\hat{y} = h(\vec{x}) \triangleq \theta^T \vec{x}$$

Equivalence #1: Given the prob. model $p(y|x,\theta)$ above the predictions are just as for our "fn. approx." linear reg. approach.

Learning

Cond. Log. Like

$$\ell(\theta) = \log \prod_{i=1}^{N} g(y^{(i)} \mid x^{(i)}, \theta, \sigma^2)$$

$$= \sum_{i=1}^{N} \log g(y^{(i)} \mid x^{(i)}, \theta, \sigma^2)$$

$$= \sum_{i=1}^{N} \left[-\log\left(\sqrt{2\sigma^2\pi}\right) - \frac{1}{2\sigma^2}\left(y^{(i)} - \theta^T x^{(i)}\right)^2 \right]$$

MCLE

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \ell(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \left[-\log\left(\sqrt{2\sigma^2\pi} - \frac{1}{2\sigma^2}\left(y^{(i)} - \theta^T x^{(i)}\right)^2 \right]$$

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

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

(the sum is the MSE!)

Equivalence #2:

MCLE for above prob. $p(y|x,\theta) = \theta^T x + \epsilon$ is equivalent to minimizing MSE for $h(\vec{x}) = \theta^T x$