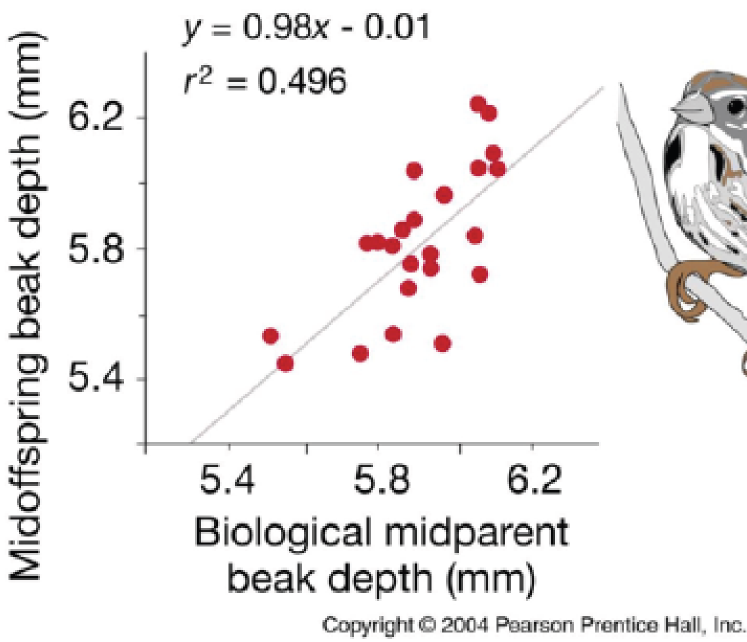
Lecture 2
Linear Regression

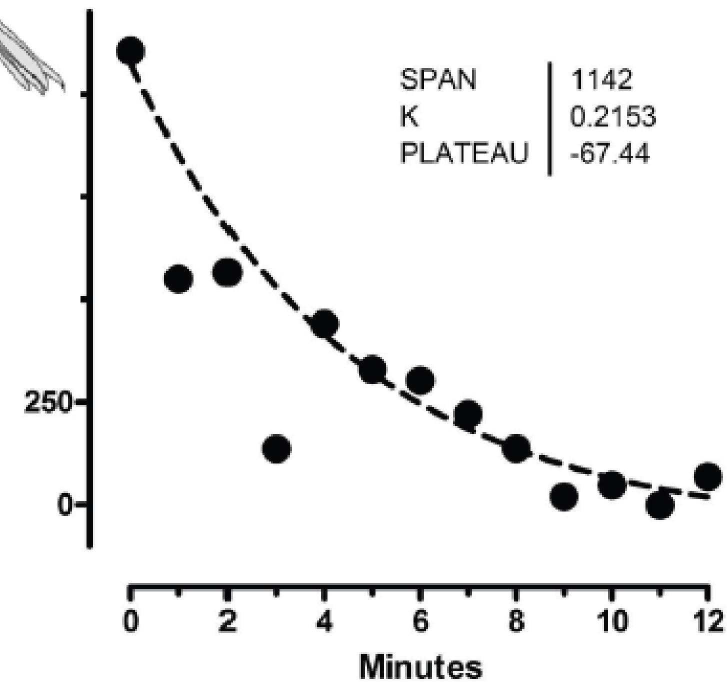
Rui Xia
School of Computer Science & Engineering
Nanjing University of Science & Technology
http://www.nustm.cn/~rxia

Regression



linear

nonlinear



Data, Input, Output, Relation

- Training data set

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

One training example $(x^{(i)}, y^{(i)})$, where i denotes the index of the example

Input: Feature Vector $\mathbf{x} = [x_1, x_2]$ Output: y

- Hypothesis: linear model

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Least Mean Square (LMS)

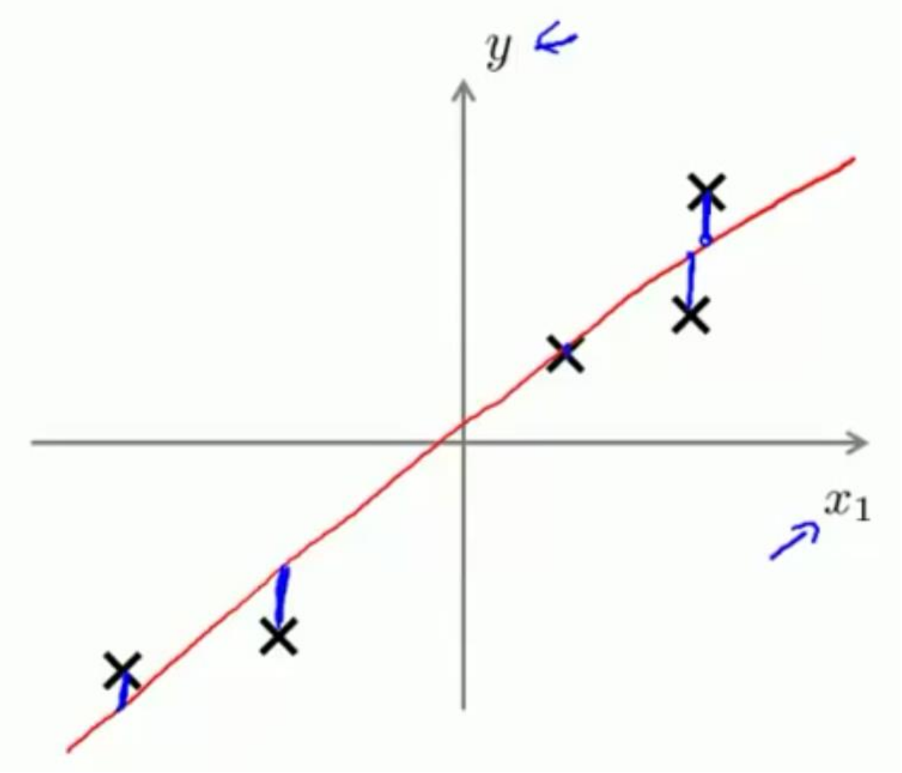
- Hypothesis

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{N} \theta_i x_i = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}$$

- Parameters $\boldsymbol{\theta}$
- Cost function

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

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



- Goal

$$\boldsymbol{\theta}^* = \arg_{\boldsymbol{\theta}} \min J_l(\boldsymbol{\theta})$$

Close-form Solution of LMS

- Define

$$\boldsymbol{X} = \begin{bmatrix} -\left(\boldsymbol{x}^{(1)}\right)^{\mathrm{T}} - \\ -\left(\boldsymbol{x}^{(2)}\right)^{\mathrm{T}} - \\ \vdots \\ -\left(\boldsymbol{x}^{(n)}\right)^{\mathrm{T}} - \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

- Then, we have

$$\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y} = \begin{bmatrix} \left(\boldsymbol{x}^{(1)}\right)^{\mathrm{T}}\boldsymbol{\theta} \\ \vdots \\ \left(\boldsymbol{x}^{(n)}\right)^{\mathrm{T}}\boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(1)}) - y^{(1)} \\ \vdots \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(n)}) - y^{(n)} \end{bmatrix}$$

- Now, the LMS cost function

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 = \frac{1}{2}(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})$$

Close-form of LMS Solution

- Calculating LMS gradient by matrix derivatives

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \frac{1}{2}(\boldsymbol{X\theta} - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{X\theta} - \boldsymbol{y}) \\ &= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta} - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\theta} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\right) \\ &= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\mathrm{tr}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta} - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\theta} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\right) \\ &= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\left(\mathrm{tr}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta} - 2\mathrm{tr}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\theta}\right) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\end{aligned}$$

- The close-form solution is obtain by letting the gradient equals zero

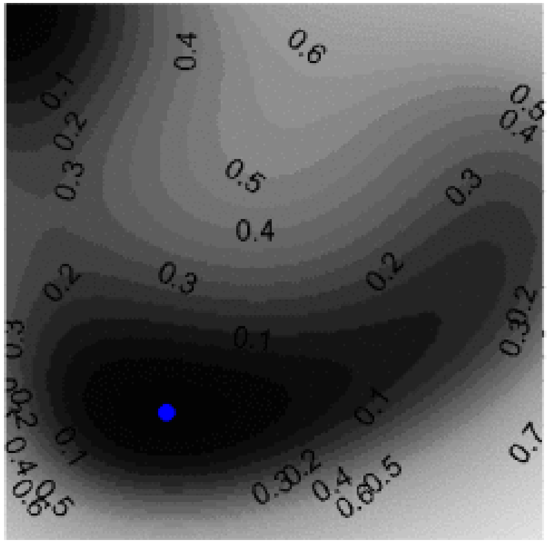
$$\boldsymbol{\theta}^{*} = \boxed{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

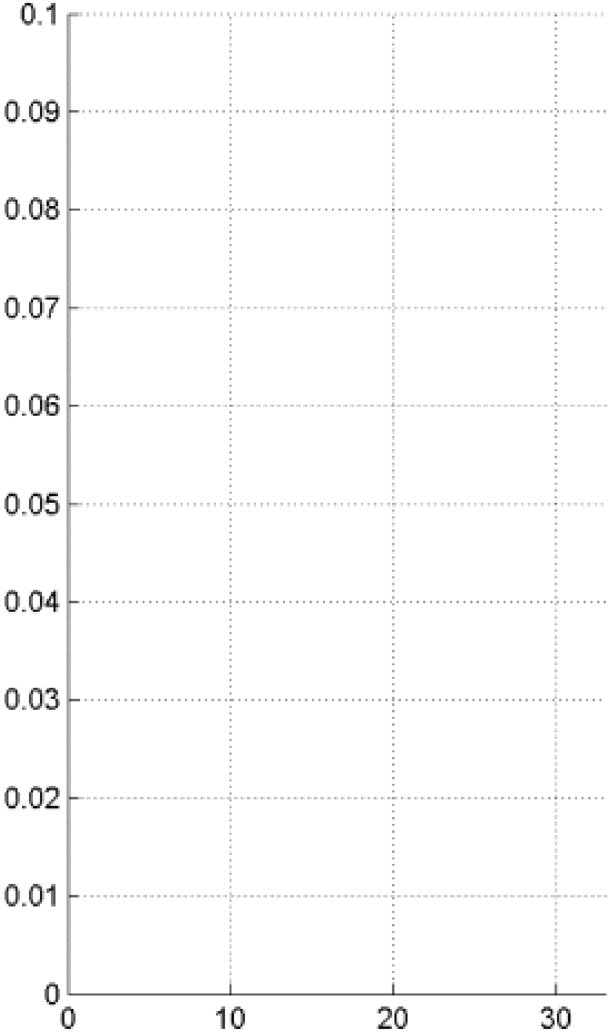
Sometimes very hard to compute!

Gradient Descent for Numeric Optimization

- Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function $f(\boldsymbol{\theta})$.
- Key idea:
  - The gradient direction is the direction that the function value increases the fastest.
- Optimization Process:
  - Start at a initial position (i.e., initial parameter $\boldsymbol{\theta}^{(0)}$)
  - At current position $\boldsymbol{\theta}^{(t)}$, repeat till convergence
    - Compute the gradient at current position: $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$
    - Move to the next position along the opposite direction of the gradient: $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha \cdot \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$, where α is the learning rate
    - $t = t + 1$

A Dynamic Illustration of Gradient Descent





Gradient Descent for Linear Regression

- Gradient

$$\frac{\partial J_l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}}\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2}\cdot 2\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\cdot\frac{\partial}{\partial \boldsymbol{\theta}}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)$$

$$= \sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\frac{\partial}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}^{(i)}\right)$$

$$= \sum_{i=1}^{N}\boxed{\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}^{(i)}}$$

"Error · Feature"

- Gradient Descent (GD) Optimization

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha\frac{\partial}{\partial \boldsymbol{\theta}}J_l(\boldsymbol{\theta}) = \boldsymbol{\theta} - \alpha\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}^{(i)}$$

Practice 1: Nanjing Housing Price Prediction

- Given history data

  Year x = [2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013]

  Price y = [2.000, 2.500, 2.900, 3.147, 4.515, 4.903, 5.365, 5.704, 6.853, 7.971, 8.561, 10.000, 11.280, 12.900]

- Assumption: the price and year are in a linear relation, thus they could be modeled by linear regression
- Task
  - To get the relationship of x and y by using linear regression, based on 1) close-form solution and 2) gradient descent;
  - To predict the Nanjing housing price in 2014.

Any Questions?