Math 7243

Machine Learning and Statistical Learning Theory

Section 2. Linear Regression

Instructor: He Wang
Department of Mathematics
Northeastern University

Review and outline:

- Supervised Learning.
  - **Regression**
  - Classification
- Unsupervised Learning.
  - Clustering Examples
  - Dimensional reduction
- Reinforcement Learning.

House Price Example:

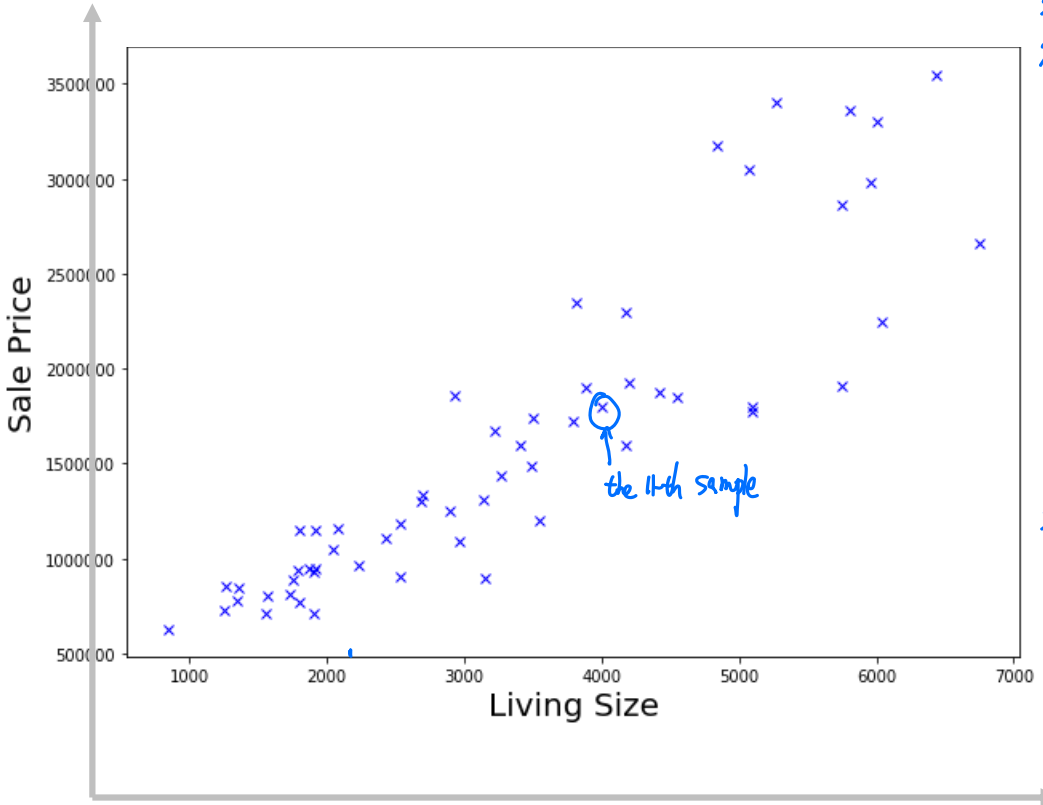
Consider the 59 single family residential houses sold in Newton, MA in Dec. 2020. (Data downloaded from www.redfin.com)

x_1	x_2	x_3	x_4	x_5	x_6	y
BEDS	**BATHS**	**LOCATION**	**SQUARE_FEET**	**LOT_SIZE**	**YEAR_BUILT**	**PRICE**
3	3	Newton	2969	15014	1967	1090000
3	2.5	Newton	1566	5582	1922	805000
4	2.5	Newton Corner	2532	6273	1953	905000
7	4.5	Newton Center	6748	26607	1902	2660000
4	4	West Newton	4200	20446	2007	1925000
4	2.5	Newton	2232	3966	1870	965000
2	1.5	Newton Corner	1344	5559	1851	775000
3	2.5	Newton	2898	12420	1943	1250000
2	2	West Newton	1729	4171	1953	815000
6	3	West Newton	3149	12616	1953	900000
5	3.5	West Newton	4000	12006	1912	1800000
4	3.5	West Newton	6430	30600	1920	3550000
4	1.5	Auburndale	1750	8222	1893	885000
2	2	Newton	840	5548	1955	630000
…	….	…	…	…	…	…

- Predict house price via living size (square_feet)

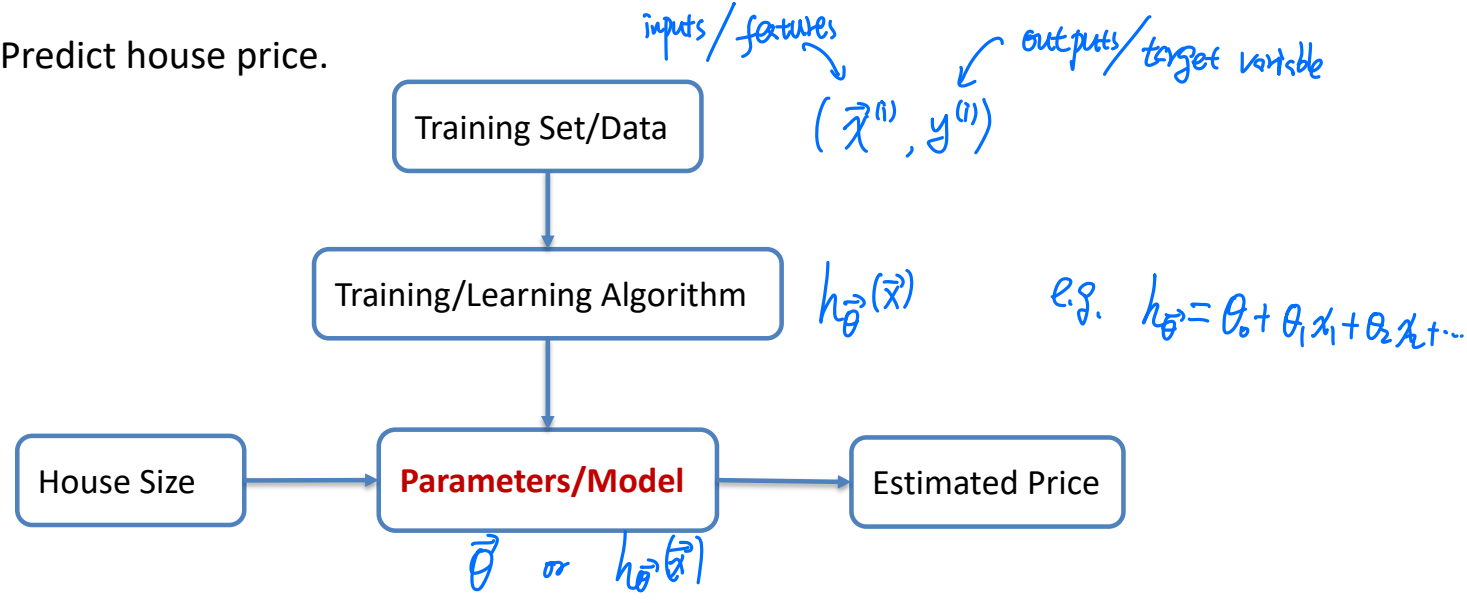
Input: a dataset that contains n samples

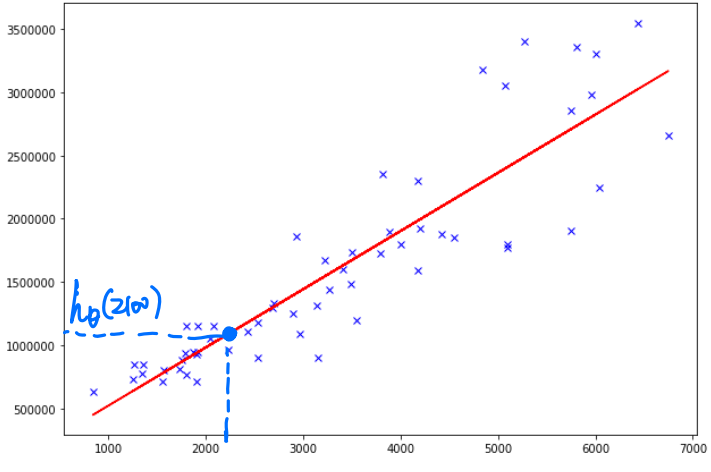
Task: if a house has x square feet, predict its price?



SQUARE_FEET	PRICE
2969	1090000
1566	805000
2532	905000
6748	2660000
4200	1925000
2232	965000
1344	775000
2898	1250000
1729	815000
3149	900000
4000	1800000
6430	3550000
1750	885000
840	630000
...	...

- Predict house price.

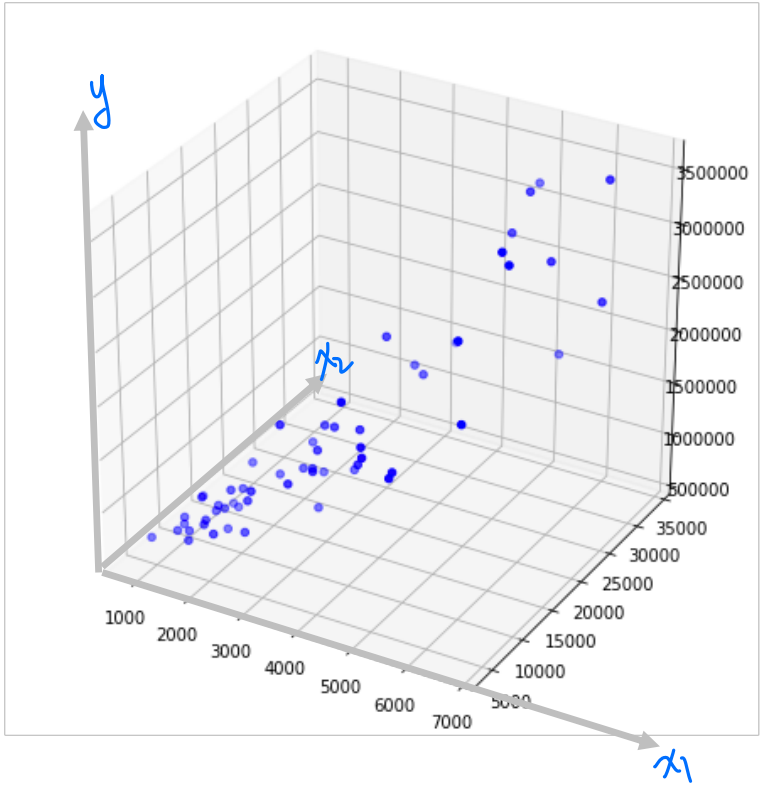




- Predict house price.

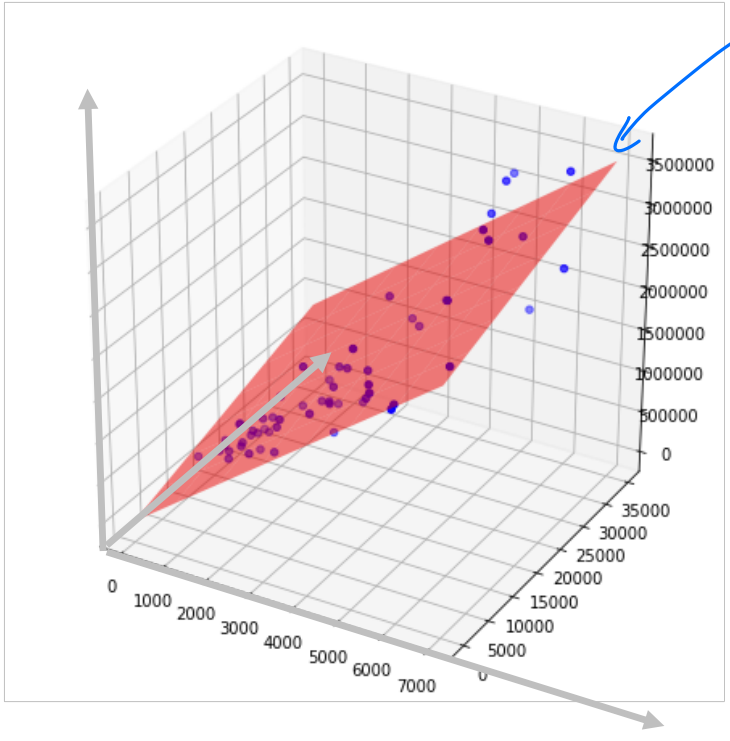
Input: a dataset that contains n samples

Task: if a house has x_1 (ft^2) living size and x_2 (ft^2) lot size, predict its price?



SQUARE_FEET (x_1)	LOT_SIZE (x_2)	PRICE (y)
2969	15014	1090000
1566	5582	805000
2532	6273	905000
6748	26607	2660000
4200	20446	1925000
2232	3966	965000
1344	5559	775000
2898	12420	1250000
1729	4171	815000
3149	12616	900000
4000	12006	1800000
6430	30600	3550000
1750	8222	885000
840	5548	630000
…	…	…

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



• Convention: vector $\vec{x}$ always in column vector.

➢ Linear Regression (Parametric Method)

Input: a dataset that contains n samples

$$D = \{(\vec{x}^{(1)}, y^{(1)}), \ldots, (\vec{x}^{(n)}, y^{(n)})\}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix} \xrightarrow{h_{\vec{\theta}}(\ )} y$$

x_0	x_1 SQUARE_FEET	x_2 LOT_SIZE	x_3 BEDS	x_4 BATHS	y PRICE
1	2969	15014	3	3	1090000
1	1566	5582	3	2.5	805000
1	2532	6273	4	2.5	905000
1	6748	26607	7	4.5	2660000
1	4200	20446	4	4	1925000
1	2232	3966	4	2.5	965000
1	1344	5559	2	1.5	775000
1	2898	12420	3	2.5	1250000
1	1729	4171	2	2	815000
1	3149	12616	6	3	900000

$\vec{x}^{(1)T}$

Assumption: linear:

$$h(\vec{x}) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$$

$$= \vec{x} \cdot \vec{\theta}$$

$$= \vec{\theta}^T \vec{x}$$

$$= \vec{x}^T \vec{\theta}$$

$$X = \begin{bmatrix} \vec{x}^{(1)T} \\ \vdots \\ \vec{x}^{(n)T} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ & \cdots & & & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$$

Goal

Data $\Rightarrow$ $h(\vec{x}^{(i)}) = y^{(i)}$ for $i = 1, 2, \cdots, n$.

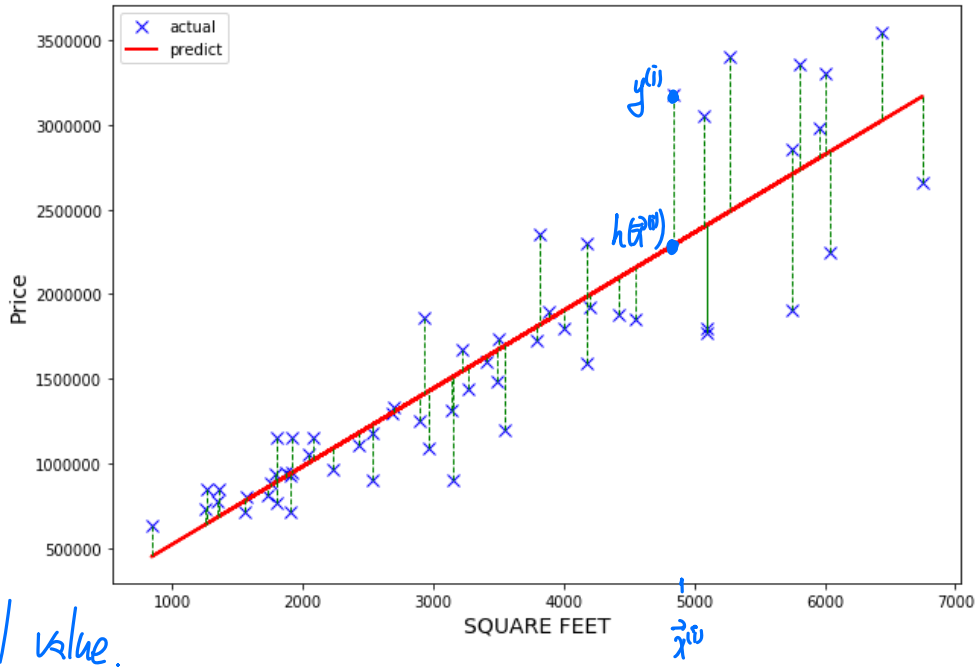
$\vec{x}^{(i)T}\vec{\theta} = y^{(i)}$

$\Rightarrow$ Solve.

$X\vec{\theta} = \vec{y}$

$$\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

- Evaluate the error:



cost vector:

$$h(X) - \vec{y} = \begin{bmatrix} h(\vec{x}^{(1)}) - y^{(1)} \\ \vdots \\ h(\vec{x}^{(n)}) - y^{(n)} \end{bmatrix}$$

Difference between predicted value and actual value.

- Review of Norm of $\mathbb{R}^n$ or any vector space

Norm is a function

$$\mathbb{R}^n \xrightarrow{\|\cdot\|} \mathbb{R}$$

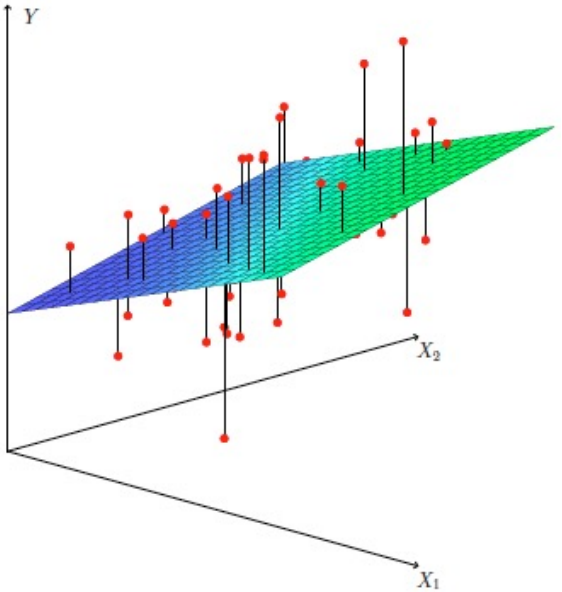
such that

① $\|\vec{x}\| \geq 0$ and $\|\vec{x}\| = 0 \iff \vec{x} = \vec{0}$

② $\|c\vec{x}\| = |c|\,\|\vec{x}\|$

③ $\|\vec{x}+\vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$

for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

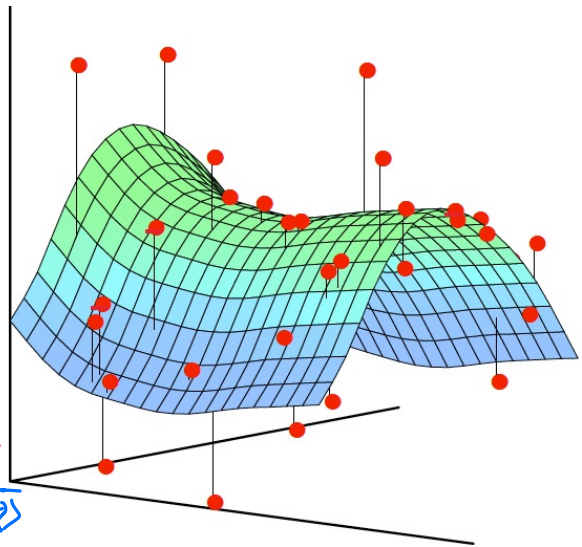


Example (l_p norm)

l_1 norm : $\|\vec{x}\|_1 = \sum_{i=1}^{n} |x_i|$

l_2 norm : $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} =: \|\vec{x}\|$ by dot product $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle}$

l_p norm : $\|\vec{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$ ← not by any inner prod.



- Cost/Loss Functions

Mean absolute error

$$L(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n}|h(\vec{x}^{(i)}) - y^{(i)}| = \frac{1}{n}\|h(X)-\vec{y}\|_1$$

or use $L(\vec{\theta}) = \frac{1}{n}\|h(X)-\vec{y}\|_p$

gradient descent to minimize $L(\vec{\theta})$

Mean Residual Sum of Squares

$$L(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(h(\vec{x}^{(i)}) - y^{(i)})^2 = \frac{1}{n}\|h(X)-\vec{y}\|^2$$

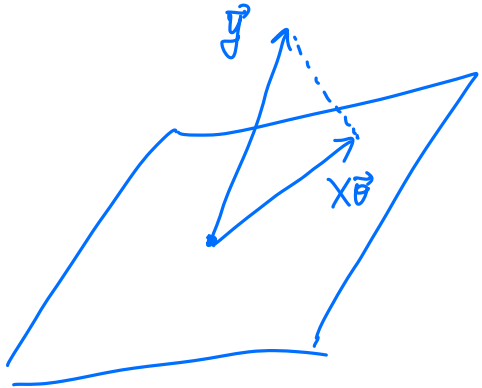
Residual Sum of Squares:

$$RSS(\vec{\theta}) = \|h(X)-\vec{y}\|^2 = \|X\vec{\theta}-\vec{y}\|^2$$

$h(X) = X\vec{\theta}$

Find $\vec{\theta}$ such that $L(\vec{\theta}) = \frac{1}{n}RSS(\vec{\theta})$ is minimized. $\Leftrightarrow$ Find $\arg\min_{\vec{\theta}} RSS(\vec{\theta})$

- Minimize Cost/Loss Functions $RSS(\vec{\theta}) = \|X\vec{\theta} - \vec{y}\|^2$



$\text{im}X = \{X\vec{\theta} \mid \forall \vec{\theta}\}$

By linear algebra

$(\text{im}X)^{\perp} = \ker(X^T)$

- To minimize the distance between $\vec{y}$ and $X\vec{\theta}$,

we need to solve $X\vec{\theta} = \text{proj}_{\text{im}X}\vec{y}$ ← only defined by inner prod.

- Equivalently, we can solve the normal equation

$$\underbrace{X^TX}_{(P+1)\times(P+1)}\vec{\theta} = X^T\vec{y}$$

Lemma: $\text{rank}\, X^TX = \text{rank}\, X$.

# of columns of X $\Rightarrow$ full column rank.

In particular, if $\operatorname{rank} X = p+1$, then $X^T X$ is invertible,

then $\vec{\theta} = (X^T X)^{-1} X^T \vec{y}$ ← easy to program.

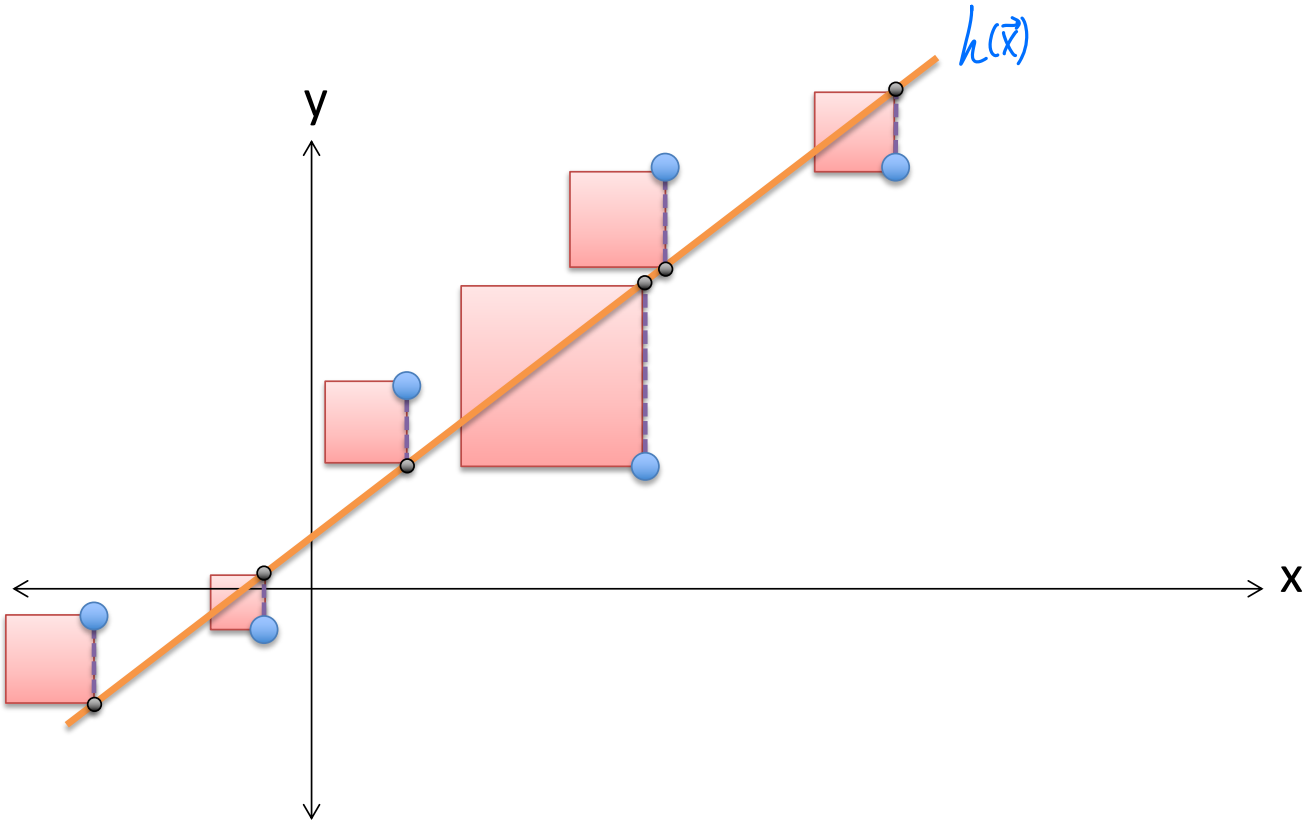
- If $\operatorname{rank} X = p+1$, and $X = QR$ $\quad Q^T Q = I_{d+1}$

  Q: orthogonal; R: upper triangular

  then $\vec{\theta} = R^{-1} Q^T \vec{y}$

Pictorial Interpretation of Squared Error

$$RSS(\vec{\theta}) = \|h(X) - \vec{y}\|^2 = \sum_{i=1}^{n} \left(h(\vec{x}^{(i)}) - y^{(i)}\right)^2$$



$$RSS(\vec{\theta}) = \|X\vec{\theta} - \vec{y}\|^2$$

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

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

$$= \vec{\theta}^T X^T X\vec{\theta} - \vec{\theta}^T X^T \vec{y} - \vec{y}^T X\vec{\theta} + \vec{y}^T\vec{y}$$

$$= \vec{\theta}^T X^T X\vec{\theta} - 2\vec{y}^T X\vec{\theta} + \vec{y}^T\vec{y}$$

↑ quadratic form　　↑ linear

$$\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

"variable vector"

$RSS(\vec{\theta})$ is a quadratic-linear function. ($\theta_i\theta_j$, θ_i, constants)

To minimize $RSS(\vec{\theta})$, we need to find critical points

- Matrix Calculus

Define:

① If $f: \mathbb{R}^n \to \mathbb{R}$, the gradient of f is

$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \vec{x} \to f(\vec{x})$

$$\frac{\partial f}{\partial \vec{x}} =: \nabla_{\vec{x}} f := \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

$\nabla_{\vec{x}} f : \mathbb{R}^n \to \mathbb{R}^n$

② If $f: \mathbb{R}^{m \times n} \to \mathbb{R}$, the gradient of f is

$\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = X \to f(X)$

$$\frac{\partial f}{\partial X} := \nabla_X f := \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & & \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

③ If $f: \mathbb{R}^n \to \mathbb{R}^m$, the derivative of $F = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix}$ is

$\vec{x} \to F(\vec{x}) = \begin{bmatrix} f_1(\vec{x}) \\ \vdots \\ f_m(\vec{x}) \end{bmatrix}$

$$\frac{\partial F}{\partial \vec{X}} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla f_1 & \cdots & \nabla f_m \end{bmatrix} \in \mathbb{R}^{n \times m}$$

· Denominator layout notation.

· Numerator layout is the transpose., that is, $\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$

Thm (linear property)

① $\nabla(f+g) = \nabla f + \nabla g$

② $\nabla(cf) = c\nabla g$ $c \in \mathbb{R}$

| True for all $\nabla_{\vec{x}}$, ∇_x, $\frac{\partial F}{\partial \vec{x}}$

Prop: If $f(\vec{x}) = \vec{b}^T \vec{x}$, then $\nabla f = \vec{b}$.

Pf: $f(x) = b_1 x_1 + \cdots + b_n x_n$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \vec{b}$$

prop: If $F(\vec{x}) = A\vec{x}$, then $\frac{\partial F}{\partial \vec{x}} = A^T$

Pf: Suppose $\vec{x} \in \mathbb{R}^n$, $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$ $\quad F: \mathbb{R}^n \to \mathbb{R}^m$

$$F(\vec{x}) = A\vec{x} = \begin{bmatrix} R_1\vec{x} \\ \vdots \\ R_m\vec{x} \end{bmatrix}$$

$$\frac{\partial F}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & & \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}_{n \times m} = \begin{bmatrix} R_1^T & \cdots & R_m^T \end{bmatrix} = \begin{bmatrix} R_1 \\ \vdots \\ R_m \end{bmatrix}^T = A^T$$

prop: (quadratic function)

If $f(\vec{x}) = \vec{x}^T A \vec{x}$, then $\nabla f = (A^T + A)\vec{x}$ where $A \in \mathbb{R}^{n \times n}$.

Pf: $f(\vec{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} a_{i1} x_i + \sum_{j=1}^{n} a_{1j} x_j \\ \vdots \\ \sum_{i=1}^{n} a_{in} x_i + \sum_{j=1}^{n} a_{nj} x_j \end{bmatrix} = A^T\vec{x} + A\vec{x}$$

Thm (product Rule)

Suppose $F: \mathbb{R}^n \to \mathbb{R}^m$ and $G: \mathbb{R}^n \to \mathbb{R}^m$ and $H = G^T F$.

$\vec{z} \to F(\vec{z})$ $\qquad$ $\vec{z} \to G(\vec{z})$

$H: \mathbb{R}^n \to \mathbb{R}$.

Then $\frac{\partial H}{\partial \vec{z}} = \frac{\partial G}{\partial \vec{z}} F + \frac{\partial F}{\partial \vec{z}} G$

$n \times 1$ $\quad$ $n \times m$ $\quad$ $m \times 1$ $\quad$ $n \times m$ $\quad$ $m \times 1$

Warning: order of product is important.

Pf: $H = G^T F = [g_1 \cdots g_m] \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix} = g_1 f_1 + \cdots + g_m f_m$

For each i

$$\frac{\partial H}{\partial z_i} = \left(\frac{\partial g_1}{\partial z_i} f_1 + \frac{\partial f_1}{\partial z_i} g_1\right) + \cdots + \left(\frac{\partial g_m}{\partial z_i} f_m + \frac{\partial f_m}{\partial z_i} g_m\right)$$

$$= \frac{\partial G}{\partial z_i} F + \frac{\partial F}{\partial z_i} G$$

$$\frac{\partial G}{\partial z_i} = \left[\frac{\partial g_1}{\partial z_i} \cdots \frac{\partial g_m}{\partial z_i}\right]$$ "row vector"

$$\frac{\partial H}{\partial \vec{z}} = \begin{bmatrix} \frac{\partial H}{\partial z_1} \\ \vdots \\ \frac{\partial H}{\partial z_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial G}{\partial z_1} F + \frac{\partial F}{\partial z_1} G \\ \vdots \\ \frac{\partial G}{\partial z_n} F + \frac{\partial F}{\partial z_n} G \end{bmatrix} = \frac{\partial G}{\partial \vec{z}} F + \frac{\partial F}{\partial \vec{z}} G$$

- Chain Rule in H.W.

Ex: $J(\vec{\theta}) = RSS(\vec{\theta}) = \vec{\theta}^T X^T X \vec{\theta} - 2\vec{y}^T X \vec{\theta} + \vec{y}^T \vec{y}$

$$\nabla_\theta J = \frac{\partial J(\vec{\theta})}{\partial \vec{\theta}} = 2X^T X \vec{\theta} - 2X^T \vec{y} = 0$$

$\Rightarrow$ Normal equation $X^T X \vec{\theta} = X^T \vec{y}$.

$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} : \mathbb{R}^n \to \mathbb{R}^n$

- Def: The Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$ is defined by

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ & - & - & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_n} & - & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$ "symmetric"

* **Thm**: If $f: \mathbb{R}^n \to \mathbb{R}$ is smooth, then a critical point $\vec{a} \in \mathbb{R}^n$

i.e. $\nabla f(\vec{a}) = \vec{0}$

is

① a local minimum if $H(f(\vec{a}))$ is positive definite. e.g. $f = x_1^2 + x_2^2$

② a ——maximum—— negative definite e.g. $f = -x_1^2 - x_2^2$

③ a saddle point if $H(f(\vec{a}))$ contains $\pm$ eigenvalues e.g. $f = x_1^2 - x_2^2$

④ For other cases, there is no conclusion. e.g. $f = x_1^3$

Prop: ① If $f(\vec{x}) = \vec{b}^T \vec{x}$, then $H(f) = 0$

② If $q(\vec{x}) = \vec{x}^T A \vec{x}$, then $H(q) = 2A$. if A is symmetric.

Ex: $H(J(\theta)) = 2X^TX$ which is positive-semidefinite.

If $\operatorname{rank} X = d+1$, then X^TX is positive-definite..