Machine Learning

Least Square

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://$\theta\theta\theta$.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

Least Square

Normal Equation:

$$\mathbf{m}\begin{bmatrix} X_0 & X_1 & X_2 & \dots & X_n \\ 1 & 100 & 2 & \dots & 3 \\ 1 & 120 & 3 & & 10 \\ . & . & . & & . \\ & & & \dots & \\ . & . & . & & . \\ . & . & . & & . \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{y} \\ \mathbf{10000\$} \\ \mathbf{10500\$} \\ . \\ . \\ . \\ . \end{bmatrix}$$

$$X = \begin{bmatrix} X_0 & X_1 & X_2 & \dots & X_n \\ 1 & 100 & 2 & \dots & 3 \\ 1 & 120 & 3 & & 10 \\ . & . & . & & . \\ & & & \dots & \\ . & . & . & & . \\ . & . & . & & . \end{bmatrix} \in R^{m*(n+1)} \qquad Y = \begin{bmatrix} \mathbf{10000\$} \\ \mathbf{10500\$} \\ . \\ . \\ . \\ . \\ . \end{bmatrix} \in R^{m*1}$$

$$y = X\,\theta \qquad (y \in R^{m*1}, X \in R^{m*(1+n)}, \theta \in R^{(n+1)*1})$$

$$\theta = X^{-1} y$$

$$X = \begin{bmatrix} . & \dots & x^{1\,T} & \dots & . \\ . & \dots & x^{2\,T} & \dots & . \\ . & \dots & x^{3\,T} & & . \\ . & . & . & & . \\ . & . & . & \dots & . \\ . & . & . & & . \end{bmatrix} \in R^{m*(n+1)} \qquad x^{i} = \begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \\ . \\ . \\ . \\ x_n^i \end{bmatrix} \in R^{(n+1)*1}$$

Normal Equation

L = $\frac{1}{2m}\sum_{i=1}^{m}(y_i - x^{(i)T}\theta)^2$ → $\underset{\theta}{\text{Min L}}$ → $\frac{d\boldsymbol{L}}{d\boldsymbol{\theta}}$ = 0

Normal Equation:

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

$(X^T \in R^{(n+1)*m}, X \in R^{m*(1+n)}, X^T X \in R^{(n+1)*(n+1)})$

$(X^T X)^{-1} \in R^{(n+1)*(n+1)}$

$y \in R^{m*1}$, $\theta^* \in R^{(n+1)*1}$

$(X^T X)^{-1} \in R^{(n+1)*(n+1)}$ → $(X^T X)^{-1}: O(n^3)$

If n=1000 → $O(10^9)$

Comparison

Gradient Descent

تکراری است.

نیاز به مقدار دهی اولیه دارد.

حجم محاسبات پایین است.

نیا به تنظیم ابرپارامترها دارد.

نیاز به Feature scaling دارد.

Normal Equation

تکراری نیست.

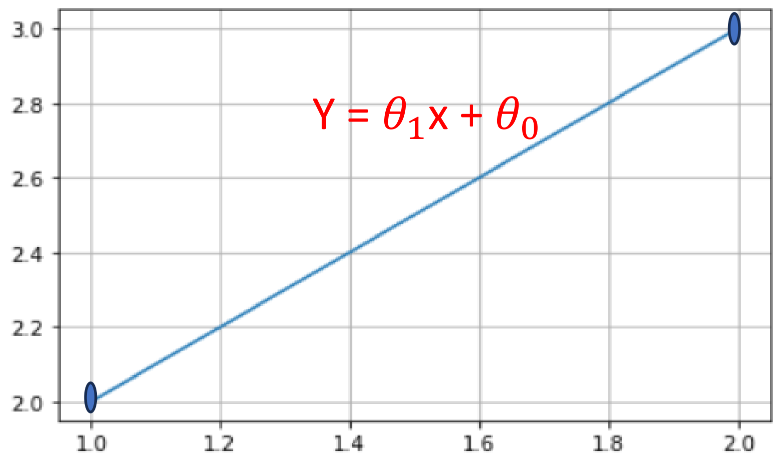
نیاز به مقداردهی اولیه پارامترها ندارد.

حجم محاسبات بالایی دارد.

نیاز به تنظیم ابرپارامترها نیست.

نیاز به Feature scaling ندارد.

یادآوری از جبر خطی

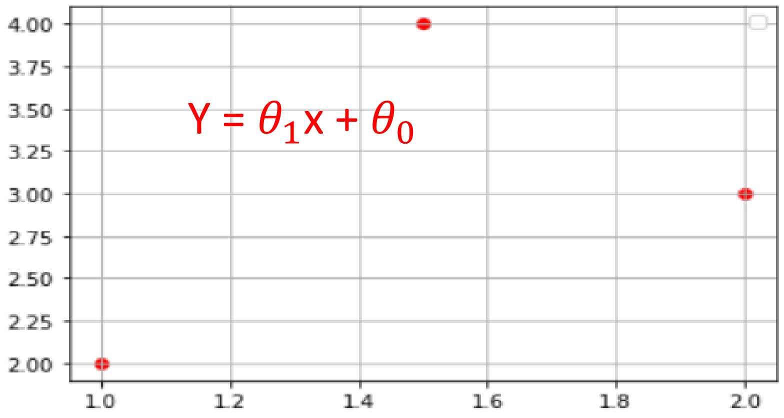


$$\begin{cases} \theta_1 + \theta_0 = 2 \\ 2\theta_1 + \theta_0 = 3 \end{cases}$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \Rightarrow \begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \theta_1 , \theta_0 = 1 \Rightarrow Y = x + 1$$

یادآوری از جبر خطی



$$
\begin{cases}
\theta_1 + \theta_0 = 2 \\
2\theta_1 + \theta_0 = 3 \\
3/2\theta_1 + \theta_0 = 4
\end{cases}
\Rightarrow
\begin{array}{c} X \\ \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3/2 & 1 \end{pmatrix} \end{array}
\begin{array}{c} \theta \\ \begin{pmatrix} \theta_1 \\ \theta_0 \end{pmatrix} \end{array}
=
\begin{array}{c} Y \\ \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} \end{array}
$$

جواب ندارد

یادآوری از جبر خطی

$$\underset{\theta_1,\theta_0}{\text{Min}}\left([2-(\boldsymbol{\theta}_1+\boldsymbol{\theta}_0)]^2+[3-(2\boldsymbol{\theta}_1+\boldsymbol{\theta}_0)]^2+[4-(3/2\boldsymbol{\theta}_1+\boldsymbol{\theta}_0)]^2\right)$$

$$\underset{\theta}{\text{Min}}\ \|y-X\boldsymbol{\theta}\|^2 \Rightarrow \begin{pmatrix} 2-(\boldsymbol{\theta}_1+\boldsymbol{\theta}_0) \\ 3-(2\boldsymbol{\theta}_1+\boldsymbol{\theta}_0) \\ 4-(3/2\boldsymbol{\theta}_1+\boldsymbol{\theta}_0) \end{pmatrix}$$

$$\mathrm{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Rightarrow \|\vec{x}\|_2=\sqrt{x_1^2+x_2^2}$$

$$x^Tx = x_1^2+x_2^2=\|x\|^2$$

مشتق گیری نسبت به بردار

$$y = a^T\boldsymbol{\theta} \qquad a = \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_0 \end{pmatrix}$$

$$y = a_1\theta_1 + a_0\theta_0 \quad \frac{d\boldsymbol{y}}{d\vec{\boldsymbol{\theta}}} = ?? \Rightarrow \frac{d\boldsymbol{y}}{d\boldsymbol{\theta}_1} = a_1, \frac{d\boldsymbol{y}}{d\boldsymbol{\theta}_0} = a_0 \Rightarrow \frac{d\boldsymbol{y}}{d\vec{\boldsymbol{\theta}}} = \begin{pmatrix} \frac{d\boldsymbol{y}}{d\boldsymbol{\theta}_1} \\ \frac{d\boldsymbol{y}}{d\boldsymbol{\theta}_2} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} = a$$

$$\boxed{\frac{da^T\theta}{d\vec{\boldsymbol{\theta}}} = a, \quad \frac{d\theta^T a}{d\vec{\boldsymbol{\theta}}} = a}$$

مشتق گیری نسبت به بردار

$$y = \boldsymbol{\theta}^T B \boldsymbol{\theta} \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \qquad \frac{d\boldsymbol{y}}{d\vec{\boldsymbol{\theta}}}$$

$$y = \begin{pmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} = \begin{pmatrix} w_1 b_{11} + w_2 b_{21} \, , \, w_1 b_{12} + w_2 b_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$$

$$= (\boldsymbol{\theta}_1 b_{11} + \boldsymbol{\theta}_2 b_{21}) \, \boldsymbol{\theta}_1 + (\boldsymbol{\theta}_1 b_{12} + \boldsymbol{\theta}_2 b_{22}) \, \boldsymbol{\theta}_2$$

$$\frac{d\boldsymbol{y}}{d\boldsymbol{\theta}_1} = 2\, \boldsymbol{\theta}_1 b_{11} + \boldsymbol{\theta}_2 b_{21} + b_{12} \boldsymbol{\theta}_2 \quad , \quad \frac{d\boldsymbol{y}}{d\boldsymbol{\theta}_2} = b_{21} \boldsymbol{\theta}_1 + \boldsymbol{\theta}_1 b_{12} + 2 \boldsymbol{\theta}_2 b_{22}$$

$$\frac{d\boldsymbol{y}}{d\vec{\boldsymbol{\theta}}} = \begin{pmatrix} \dfrac{d\boldsymbol{y}}{d\boldsymbol{\theta}_1} \\ \dfrac{d\boldsymbol{y}}{d\boldsymbol{\theta}_2} \end{pmatrix} = \left(\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \end{pmatrix} \right) \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} = \begin{pmatrix} 2b_{11} & b_{12} + b_{21} \\ b_{21} + b_{12} & 2b_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} = \boxed{(B + B^T)\, \theta}$$

مشتق گیری نسبت به بردار

If B symmetric $\Rightarrow$ B = B^T $\Rightarrow$ $\frac{dy}{d\vec{\boldsymbol{\theta}}}$ = 2Bθ

$\frac{dy}{db_{11}} = \theta_1^2$, $\frac{dy}{db_{12}} = \theta_1\theta_2$, $\frac{dy}{db_{21}} = \theta_2\,\theta_1$, $\frac{dy}{db_{22}} = \theta_2^2$

$$\frac{dy}{dB} = \theta\theta^T$$

$x^T . x = \|x\|^2$ $\Longrightarrow$ $J(\theta) = \|y - X\boldsymbol{\theta}\|^2 = (y - X\boldsymbol{\theta})^T (y - X\boldsymbol{\theta})$

اسکالر ^{T}= اسکالر

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

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

$J(\theta) = y^T y - \underbrace{y^T X \theta}_{\text{اسکالر}} - \underbrace{(X\theta)^T y}_{\text{اسکالر}} + (X\theta)^T X\theta$

اسکالر

$= y^T y - 2\, \overbrace{y^T X}^{= a^T} \theta + \theta^T \overbrace{X^T X}^{= B} \theta$

$B = X^T X$

$(X^T X)^T = X^T X$

یک ماتریس متقارن

$B = B^T$

$y^T X = a^T \longrightarrow X^T y = a$

$$\boxed{\frac{d\, y^T X \theta}{d\vec{\boldsymbol{\theta}}} = \frac{d a^T \theta}{d\vec{\boldsymbol{\theta}}} = a = X^T y}$$

$$\min_{\theta} j(\theta) = -2\, X^T y + 2B\theta = -2\, X^T y + 2\, X^T X\, \theta = 0$$

$$\cancel{2}\, X^T X\, \theta = \cancel{2}\, X^T y \Longrightarrow \theta = (X^T X)^{-1}\, X^T y = X^+ y$$

(pseudo inverse)

یادآوری از جبر خطی

بردارهایی را مستقل خطی میگویند که هیچکدام از آنها را به صورت ترکیب خطی از یکدیگر نتوان نوشت.

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \checkmark \qquad \begin{pmatrix} 1 & 2 & 8 \\ 2 & 7 & 25 \\ 3 & 4 & 18 \end{pmatrix} \times$$

$x \in R^{m*D}$

Span(row vectors): مجموعه همه ترکیب های خطی سطرها (Row space)

Span(column vectors): مجموعه همه ترکیب های خطی ستون ها (Column space)

یادآوری از جبر خطی

Null space(X) = N(x): همه بردار های غیر صفر θ که $X\theta$ برابر صفر است.

Dimension of column space: تعداد ستون های مستقل خطی

Dimension of row space: تعداد سطرهای مستقل خطی

Dim (row space) = Dim (column space)

Nullity (A) = Dim(N(A))

Rank(X) = Dim(row space)=Dim(column space)

Row Echelon Form

شکل سطری پلکانی

همه سطرهای کامل صفر در پایین ماتریس باشد.

عنصر پیشرو (pivot) هر سطر غیر صفر در سمت راست عنصر پیشرو سطر بالای آن باشد.

$$\begin{pmatrix} 1 & a_0 & a_1 & a_2 & a_3 \\ 0 & 0 & 2 & a_4 & a_5 \\ 0 & 0 & 0 & 1 & a_6 \end{pmatrix}$$

تبدیل به Row echelon form(عملیات سطری مقدماتی):

1. جابجایی دو سطر 2. ضرب عدد حقیقی در یک سطر 3. جمع یا تفریق مضربی از یک سطر با سطر دیگر

مثال

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 \end{pmatrix} \xrightarrow{-2\,R_1 + R_2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix}$$

تعداد سطرهای غیر صفر در شکل سطری پلکانی تعداد سطرهای مستقل خطی هستند.

رتبه یا Rank ماتریس تعداد سطرهای غیر صفر در شکل سطری پلکانی است.

Rank(X) + Nullity(X) = number of columns of X

مثال

$$X = \begin{pmatrix} -1 & 0 & -1 & 2 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 1 & -1 \end{pmatrix} \xrightarrow{2R_1 + R_2} \begin{pmatrix} -1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 4 \\ 1 & 0 & 1 & -1 \end{pmatrix}$$

$$\xrightarrow{R_1 + R_3} \begin{pmatrix} -1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{-4R_3 + R_2} \begin{pmatrix} -1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{\text{جابجایی سطر سه با سطر دو}} \begin{pmatrix} -1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Rank(X) = 2

Rank(X) + Nullity(X) = 4

2 + 2 =4

مثال

Dim(column space B) = ?

$$B = \begin{pmatrix} 1 & 0 & -1 & 2 \\ 2 & 1 & 2 & 3 \\ -1 & 0 & 1 & -2 \end{pmatrix}$$

Column space B = Row space B^T

$$B^T = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ -1 & 2 & 1 \\ 2 & 3 & -2 \end{pmatrix} \xrightarrow{R_1 + R_3} \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 4 & 0 \\ 2 & 3 & -2 \end{pmatrix} \xrightarrow{-2R_1 + R_4} \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 4 & 0 \\ 0 & -1 & 0 \end{pmatrix}$$

$$\xrightarrow{-4R_2 + R_3} \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \xrightarrow{R_2 + R_4} \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Dim(Column space B) = Dim(Row space B^T) = 2

$X \in R^{m*D}$

$X^T X \in R^{D*D}$ (Gram Matrix)

ماتریسی معکوس پذیر است که مربعی بوده و رتبه آن کامل باشد.

$X^T X$ در صورتی معکوس دارد که X دارای رتبه ستونی کامل باشد. (Full column rank)

به عبارت دیگر Rank(X) = D

اثبات

فرض کنید Rank(X) < D باشد با توجه به Rank(X) + N(X) = D یک u وجود دارد که Xu = 0
در نتیجه داریم $X^T X\, u = 0$ و $Rank(X^T X) < D$ و معکوس پذیر نیست.

به طور برعکس اگر $X^T X$ معکوس پذیر نباشد یک بردار v وجود دارد که $X^T X\, v = 0$

پس داریم:

$$0 = v^T\ X^T X\, v = (Xv)^T (Xv) = ||Xv||^2$$

پس Xv=0یعنی Rank(X) < D.

اگر m<D باشد در نتیجه Rank(X) < D و ماتریس $X^T X$ معکوس پذیر نیست.

ممکن است حتی وقتی m >= D باشد به دلیل وابستگی ها بین ستون های X رتبه X کمتر از D باشد.