train data : $(x[1], y[1]), (x[2], y[2]), \ldots (x[n], y[n])$

Model : $Y \approx \beta_0 + \beta_1 X$

loss function in this example was least square solution:

$$L(\beta) = \frac{1}{2} \sum_{i=1}^{n} \left(y[i] - (\beta_0 + \beta_1 x[i])\right)^2$$

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\arg\min}\, L(\beta)$$

روشهای مختلفی برای حل این معادله وجود دارد که یکی از آنها Gradient descent است. یک روش تکراری

است یعنی از یک حدس اولیه برای مقادیر پارامترها شروع می‌کنیم و در هر تکرار مقادیر را آپدیت می‌کنیم.

$$\beta_0 := \beta_0 - \underbrace{\alpha}_{\text{learning rate}} \frac{\partial}{\partial \beta_0} L(\beta) \quad ; \quad \beta_1 := \beta_1 - \alpha \frac{\partial}{\partial \beta_1} L(\beta)$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \frac{1}{\cancel{2}} \sum_{i=1}^{n} \cancel{2} \left(y[i] - \beta_0 - \beta_1 x[i]\right)(-1) = -\sum_{i=1}^{n} \left(y[i] - \beta_0 - \beta_1 x[i]\right)$$

$$\frac{\partial L(\beta)}{\partial \beta_1} = \frac{1}{\cancel{2}} \sum_{i=1}^{n} \cancel{2} \left(y[i] - \beta_0 - \beta_1 x[i]\right)(-x[i]) = -\sum_{i=1}^{n} x[i] \left(y[i] - \beta_0 - \beta_1 x[i]\right)$$

بنابراین داریم:

$$\beta_0 := \beta_0 + \alpha \sum_{i=1}^{n} \left(y[i] - \beta_0 - \beta_1 x[i]\right)$$

→ batch gradient descent

$$\beta_1 := \beta_1 + \alpha \sum_{i=1}^{n} \left(x[i]\left(y[i] - \beta_0 - \beta_1 x[i]\right)\right)$$

Subject:

Year. Month. Date. ( )

در واقع به این حالت که از تمام training data استفاده کردیم برای آپدیت کردن پارامترها می‌گوییم batch gradient descent.

در یک روش دیگر به جای در نظر گرفتن تمام نمونه‌ها تنها یک observation را در نظر می‌گیریم، می‌گوییم ← stocastic gradient descent

$$\beta_0 := \beta_0 + \alpha\,(y[i] - \beta_0 - \beta_1 x[i]) \qquad \text{for } i=1,\dots,n$$

$$\beta_1 := \beta_1 + \alpha\,(x[i]\,(y[i] - \beta_0 - \beta_1 x[i]))$$

در واقع هزینه و زمان stocastic gradient descent کمتر از batch gradient desecent است.

⇒ linear regression analytical solution:

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y[i] - \beta_0 - \beta_1 x[i])^2$$

$$\hat{\beta} = \underset{\beta}{\arg\min}\left(\frac{1}{2}\sum_{i=1}^{n}(y[i] - \beta_0 - \beta_1 x[i])^2\right)$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = 0 \quad ; \quad \frac{\partial L(\beta)}{\partial \beta_1} = 0$$

Solve for $\beta_0, \beta_1 \Rightarrow \hat{\beta}_0, \hat{\beta}_1$

$$\frac{\partial L(\beta)}{\partial \beta_0} = 0 \Rightarrow -\sum_{i=1}^{n} \left(y[i] - \beta_0 - \beta_1 x[i]\right) = 0$$

$$\Rightarrow -\frac{1}{n}\sum_{i=1}^{n} y[i] + \beta_0 \frac{1}{n}\sum_{i=1}^{n} 1 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x[i] = 0$$

$$\Rightarrow -\bar{y} + \beta_0 + \beta_1 \bar{x} = 0 \Rightarrow \boxed{\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}}\ *$$

$$\frac{\partial L(\beta)}{\partial \beta_1} = 1 \Rightarrow -\sum_{i=1}^{n} x[i]\left(y[i] - \beta_0 - \beta_1 x[i]\right) = 0$$

$$\Rightarrow -\frac{1}{n}\sum_{i=1}^{n} x[i]\,y[i] + \beta_0 \frac{1}{n}\sum_{i=1}^{n} x[i] + \beta_1 \frac{1}{n}\sum_{i=1}^{n} (x[i])^2$$

$$\Rightarrow -\overline{xy} + \beta_0 \bar{x} + \beta_1 \overline{x^2} = 0$$

$$* \overset{\text{جایگذاری}}{\Rightarrow} -\overline{xy} + \left(\bar{y}\bar{x} - \beta_1 (\bar{x})^2\right) + \beta_1 \overline{x^2} = 0$$

$$\Rightarrow \boxed{\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}}$$

برای اینکه ثابت کنیم این نقاط مینیمم تابع هستند با شرط convex بودن تابع باید مشتق دوم مثبت باشد پس باید مشتق دوم را حساب کنیم:

$$\frac{\partial^2 L(\beta)}{\partial \beta_0^2} = +\sum_{i=1}^{n} 1 = n > 0$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_1^2} = +\sum_{i=1}^{n} (x[i])^2 > 0$$

$\Rightarrow$ نقاط بدست آمده مینیمم تابع هستند.

Subject:

Year. Month. Date. ( )

Example 1.

i	$x[i]$	$y[i]$	$x[i]y[i]$	$(x[i])^2$
۱	۱	۴,۸	۴,۸	۱
۲	۳	۱۱,۳	۳۳,۹	۹
۳	۵	۱۷,۲	۸۶	۲۵
$\frac{1}{n}\sum_{i=1}^{n}$	۳	۱۱,۱	۴۱,۵۷	۱۱,۶۷

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{۴۱,۵۷ - (۳)(۱۱,۱)}{(۱۱,۶۷) - (۹)} = \boxed{۳,۱}$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = ۱۱,۱ - (۳,۱)(۳) = \boxed{۱,۸}$$

if $x = ۴ \Rightarrow$ Predict y ? $\quad \hat{y} = ۱,۸ + (۳,۱)(۴) = \boxed{۱۴,۲}$

← در simple LR همانطور که گفته شد تعداد متغیرهای explanatory یک بود اما

در multiple linear regression تعداد متغیرهای explanatory بیشتر از یک عدد است.

$X_1, X_2, \dots, X_p \longrightarrow Y$

$$Y \approx \beta_0 X_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad ; \quad X_0 = 1 \Rightarrow Y \approx \sum_{j=0}^{p} \beta_j X_j$$

want to estimate $\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_p$

Example 2 :

i	x_0	x_1	$x_۲$	y
۱	۱	۳	۱	۷
۲	۱	۱	-۱	-۲
۳	۱	۷	-۲	۲

$$x[۱] = \begin{pmatrix} ۱ \\ ۳ \\ ۱ \end{pmatrix} = [۱ \;\; ۳ \;\; ۱]^T$$

$$x[۲] = \begin{pmatrix} ۱ \\ ۱ \\ -۱ \end{pmatrix} = [۱ \;\; ۱ \;\; -۱]^T$$

$$x[۳] = \begin{pmatrix} ۱ \\ ۷ \\ -۲ \end{pmatrix} = (۱ \;\; ۷ \;\; -۲)^T$$

$$y = \begin{pmatrix} ۷ \\ -۲ \\ ۲ \end{pmatrix}$$

حالت کلی داریم:

$$x[i] = \begin{pmatrix} x_0[i] \\ x_1[i] \\ \vdots \\ x_p[i] \end{pmatrix}$$

$$\beta = (\beta_0 \;\; \beta_1 \;\; \dots \;\; \beta_p)^T$$

$$\hat{y}[i] = \sum_{j=0}^{p} \beta_j x_j[i] = \beta^T x[i]$$

$$L(\beta) = \frac{1}{۲} \sum_{i=1}^{n} (y[i] - \beta^T x[i])^2$$

$$X = \underbrace{\begin{bmatrix} 1 & x_1[1] & \dots & x_p[1] \\ 1 & x_1[2] & \dots & x_p[۲] \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1[n] & \dots & x_p[n] \end{bmatrix}}_{n \times (p+1)} \; ; \; \beta = \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{(p+1) \times 1}$$

$$\Rightarrow X\beta = \underbrace{\begin{bmatrix} x[1]^T \beta \\ x[2]^T \beta \\ \vdots \\ x[n]^T \beta \end{bmatrix}}_{n \times 1} = \begin{bmatrix} \hat{y}[1] \\ \hat{y}[2] \\ \vdots \\ \hat{y}[n] \end{bmatrix}$$

$$L(\beta) = \frac{1}{2} \underbrace{(Y - X\beta)^T} \underbrace{(Y - X\beta)}$$

$$\begin{bmatrix} y[1]-\hat{y}[1] & y[2]-\hat{y}[2] & \cdots & y[n]-\hat{y}[n] \end{bmatrix} \qquad \begin{bmatrix} y[1] \\ y[2] \\ \vdots \\ y[n] \end{bmatrix} - \begin{bmatrix} x[1]^T\beta \\ x[2]^T\beta \\ \vdots \\ x[n]^T\beta \end{bmatrix} = \begin{bmatrix} y[1]-\hat{y}[1] \\ y[2]-\hat{y}[2] \\ \vdots \\ y[n]-\hat{y}[n] \end{bmatrix}$$

$$= \sum_{i=1}^{n} (y[i] - \hat{y}[i])^2$$

$$\hat{\beta} = \arg\min_{\beta} L(\beta)$$

$$\nabla_{\beta} L(\beta) = \frac{\partial L(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_p} \end{bmatrix} \quad \Rightarrow \beta := \beta - \alpha \nabla_{\beta} L(\beta)$$

برای اینکه بردار β بدست آمده min باشد باید مشتق دوم تابع مثبت باشد در اینجا مشتق دوم $L(B)$ را هسین می نامیم و چون یک ماتریس است برای آن تعریف معین مثبت را داریم.

$$\nabla^2_{\beta} L(\beta) = \begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0^2} & \frac{\partial L(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial L(\beta)}{\partial \beta_0 \partial \beta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial L(\beta)}{\partial \beta_p \partial \beta_0} & \frac{\partial L(\beta)}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial L(\beta)}{\partial \beta_p^2} \end{bmatrix}$$

Subject:

Year. Month. Date. ( )

Example 3. $f(x) = x_1^2 + 3x_2^2$; $\nabla_x f(x) = ?$ $\nabla_x^2 f(x) = ?$

$$\nabla f_x(x) = \begin{pmatrix} \dfrac{\partial f(x)}{\partial x_1} \\ \dfrac{\partial f(x)}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 6x_2 \end{pmatrix} \; ; \; \nabla_x^2 f(x) = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$$

Example 4. $f(x) = x_1^3 x_2 + 2x_2$

$$\nabla_x f(x) = \begin{pmatrix} 3x_1^2 x_2 \\ x_1^3 + 2 \end{pmatrix} \qquad \nabla_x^2 f(x) = \begin{pmatrix} 6x_1x_2 & 3x_1^2 \\ 3x_1^2 & 0 \end{pmatrix}$$

Example 5. $f(\beta) = \underset{(p+1)\times 1}{X^T} \underset{(p+1)\times 1}{\beta}$

$$f(\beta) = \sum_{i=0}^{p} x_i \beta_i = \beta_0 x_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

$$\nabla_\beta f(\beta) = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \; ; \; \nabla_\beta^2 f(\beta) = 0_{(p+1)(p+1)}$$

Example 6. $f(\beta) = \beta^T \beta \Rightarrow f(\beta) = \beta_0^2 + \beta_1^2 + \dots + \beta_p^2$

$$\nabla_\beta f(\beta) = \begin{pmatrix} 2\beta_0 \\ 2\beta_1 \\ \vdots \\ 2\beta_p \end{pmatrix} \; ; \; \nabla_\beta^2 f(\beta) = \begin{pmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 2 \end{pmatrix} = 2I_{(p+1)(p+1)}$$