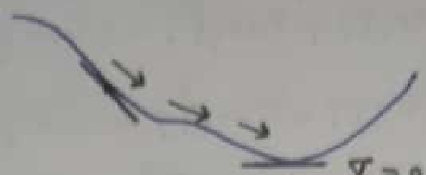
۲) الگوریتم Gradient Descent به صورت iterative عمل می‌کند و با استفاده از مشتق نقطه می‌کوشد تا به نقطه minimum برسد. یعنی برخلاف gradient تابع در نقطه کنونی سعی می‌کند که به minimum برسد.



$\nabla = 0 \Rightarrow$ minimum

$$\frac{\partial}{\partial w} J(w,b) = \frac{\partial}{\partial w} \frac{1}{2} (h(x) - y)^2 = (h(x) - y) \frac{\partial}{\partial w} (h(x) - y)$$

$$= (h(x) - y) \frac{\partial}{\partial w} \left(\frac{e^{wx+b}}{1+e^{wx+b}} \right)$$

$$= (h(x) - y) \left(\frac{x e^{(wx+b)} (1+e^{(wx+b)} - e^{wx+b})}{(1+e^{wx+b})^2} \right)$$

$$= (h(x) - y) \left(\frac{x e^{wx+b}}{1+e^{wx+b}} \right) \left(1 - \frac{e^{wx+b}}{1+e^{wx+b}} \right)$$

$$= x (h(x) - y) \left(\frac{e^{wx+b}}{1+e^{wx+b}} \right) \left(1 - \frac{e^{wx+b}}{1+e^{wx+b}} \right)$$

$$= x (h(x) - y) (h(x)) (1 - h(x))$$

$$\frac{\partial}{\partial b} J(w,b) = \frac{\partial}{\partial b} \frac{1}{2} (h(x) - y)^2 = (h(x) - y) \frac{\partial}{\partial b} (h(x) - y)$$

$$= (h(x) - y) \frac{\partial}{\partial b} \left(\frac{e^{wx+b}}{1+e^{wx+b}} \right)$$

$$= (h(x) - y) \left(\frac{e^{wx+b}}{(e^{wx+b}+1)^2} \right) = (h(x) - y)(h(x))(1 - h(x))$$

α: learning rate

$$w := w - \alpha \frac{\partial}{\partial w} J(w,b)$$

$$b := b - \alpha \frac{\partial}{\partial b} J(w,b)$$

۴

۳)

$$\hat{Y} = \alpha + \beta X$$

$$E[\hat{Y}] = E[\alpha + \beta X] = \beta E[X] + \alpha = \beta \mu_x + \alpha$$

$$E[Y - \hat{Y}] = E[Y] - E[\hat{Y}] = \mu_Y - \beta \mu_x - \alpha$$

$$Var(Y - \hat{Y}) = E[(Y - \hat{Y})^2] - (E[Y - \hat{Y}])^2$$

$$= Var(Y) + Var(\hat{Y}) - 2Cov(Y, \hat{Y})$$

$$= Var(Y) + Var(\hat{Y}) - 2Cov(Y, \alpha + \beta X)$$

$$= Var(Y) + Var(\hat{Y}) - 2Cov(Y, X)$$

$$Var(\hat{Y}) = Var(\alpha + \beta X) = \beta^2 Var(X) = \beta^2 \delta_x^2$$

$$\rightarrow Var(Y - \hat{Y}) = \delta_y^2 + \beta^2 \delta_x^2 - 2\beta Cov(X, Y)$$

$$= \delta_y^2 + \beta^2 \delta_x^2 - 2\beta \delta_{XY}$$

$$\Rightarrow E[(Y - \hat{Y})^2] = \beta^2 \delta_x^2 + \delta_y^2 - 2\beta \delta_{xy} + (\mu_y - \beta \mu_x - \alpha)^2$$

minimize $\Rightarrow \mu_y - \beta \mu_x - \alpha = 0 \rightarrow \alpha = \boxed{\mu_y - \beta \mu_x}$

حال که تکلیف α مشخص شده است با جایگذاری α، عبارت $E[(Y-\hat{Y})^2]$ به یک تابع با آرگومان β تبدیل می‌شود. (با داده ثابت μ_x و μ_y و ... همگی شش ثابت هستند) پس برای رسیدن به β برای بهینه‌سازی از عبارت زیر مشتق می‌گیریم:

$$E[(Y - \hat{Y})^2] = \beta^2 \delta_x^2 + \delta_y^2 - 2\beta \delta_{xy}$$

$$\xrightarrow{\partial/\partial\beta = 0} 2\beta \delta_x^2 - 2\delta_{xy} = 0 \rightarrow \boxed{\beta = \frac{\cancel{2}\delta_{xy}}{\cancel{2}\delta_x^2} = \frac{\delta_{xy}}{\delta_x^2}}$$

$$\frac{Var(Y) - Var(Y - \hat{Y})}{Var(Y)} = \frac{\cancel{\delta_y^2} - \cancel{\delta_y^2} - \beta^2 \delta_x^2 + 2\beta \delta_{xy}}{\delta_y^2} \xrightarrow{\text{جایگذاری } \beta} \frac{\frac{-\delta_{xy}^2}{(\delta_x^2)^{\cancel{2}}} \cancel{\delta_x^2} + \frac{2\delta_{xy}^2}{\delta_x^2}}{\delta_y^2}$$

$$\rightarrow = \frac{\delta_{xy}^2}{\delta_x^2 \delta_y^2} = r_{xy}^2$$

۳

۴)

a

* Least squares estimation

$$Y = f(x) + \text{noise} \xrightarrow[\text{squares}]{\text{least}} \sum_{i=1}^{n} (y_i - f_b(x_i))^2 : \| y - Xb \|^2$$

$$f(x) = \beta_0 x + \beta_1 x^2$$

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{1,n} & x_{n,2} \end{pmatrix} = (x_1 \quad x_2) , n = k+1$$

باتوجه به بحث های انجام شده سر کلاس درباره با اثبات Maximum Likelihood

می‌توان گفت ، (رابطه نرمال)

$$\hat{\beta} = (X'X)^{-1} X'y$$

پس به این ترتیب به تعیین پارامترها می‌رسیم .

b

ماتریس covariance برابر است با :

$$(X'X)^{-1} \delta^2$$

که δ^2 در اینجا واریانس noise ها می‌باشد که برابر است با :

(یک estimator برای آنها تعیین می‌کنیم)

$$\hat{\delta}^2 = \frac{1}{n-p} \| y - X\hat{\beta} \|^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{e}_i^2$$

$n = k+1$, $p = 2$

که $\hat{e}_i$ ها برابراند با :

$$\hat{e}_i = y_i - \hat{f}(x_i)$$

حال به اثبات این رابطه می‌پردازیم .

Σ

if
1. $Y = \beta X + \varepsilon$
2. $E[\varepsilon | X] = 0$
3. $Cov(\varepsilon | X) = \delta^2 I = V(\varepsilon | X)$
4. estimator is unique (normal equation)

then $Cov(\beta | X) = \delta^2 (X'X)^{-1}$

proof: $\beta = (X'X)^{-1} X'Y = (X'X)^{-1} X'(X\beta + \varepsilon)$

$$= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'\varepsilon$$

$$= \beta + (X'X)^{-1} X'\varepsilon$$

Taking expectation of both sides gives you: estimator is unbiased

Taking variances of both sides gives you:

$$Cov(\beta) = Cov(\beta | X) = Cov(\beta + (X'X)^{-1} X'\varepsilon | X)$$

$$= Cov((X'X)^{-1} X'\varepsilon | X)$$

$$= (X'X)^{-1} X' V(\varepsilon | \cancel{X}) X (X'X)^{-1}$$

$$= (X'X)^{-1} X' \delta^2 I X (X'X)^{-1}$$

$$= \delta^2 (X'X)^{-1} \overbrace{X' X (X'X)^{-1}}^{I}$$

$$= \delta^2 (X'X)^{-1}$$

C

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\rightarrow MSE = \frac{1}{k+1} \sum_{i=1}^{k+1} (Y_i - (\beta X + \varepsilon))$$

$$= \frac{1}{k+1} \sum_{i=1}^{k+1} (Y_i - ((X'X)^{-1} X'Y_i X + \varepsilon))$$

ه) در اینجا دقیقاً از روابط سوال ۴ استفاده خواهیم کرد :

$$\ll y = \beta_1 x + \beta_0 \gg$$

بردار y

x	y
4	31
9	58
10	65
14	73
4	37
7	44
12	60
22	91
1	21
17	84

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$X = \begin{pmatrix} 1 & 4 \\ 1 & 9 \\ 1 & 10 \\ 1 & 14 \\ 1 & 4 \\ 1 & 7 \\ 1 & 12 \\ 1 & 22 \\ 1 & 1 \\ 1 & 17 \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y \cong \begin{pmatrix} 21.69 \\ 3.47 \end{pmatrix} \begin{array}{l} \rightarrow \beta_0 \\ \rightarrow \beta_1 \end{array}$$

$$\varepsilon = y - \hat{y} = -(\beta_1 x + \beta_0) + y = \begin{pmatrix} 31 - 35.57 \\ 58 - 52.92 \\ 65 - 56.4 \\ 73 - 70.28 \\ 37 - 35.57 \\ 44 - 45.98 \\ 60 - 63.34 \\ 91 - 98.04 \\ 21 - 25.16 \\ 84 - 80.69 \end{pmatrix} = \begin{pmatrix} -4.57 \\ 5.07 \\ 8.6 \\ 2.71 \\ 1.42 \\ -1.98 \\ -3.34 \\ -7.04 \\ -4.16 \\ 3.30 \end{pmatrix}$$

$$\underset{\text{ماتریس covariance}}{\hat{\sigma}^2} = (X'X)^{-1}\hat{\varepsilon}^2 \Rightarrow \hat{\varepsilon}^2 = \frac{1}{10-2}\sum_{i=1}^{10}\hat{e}_i^2 \cong \frac{1}{8}(223.04) \cong 27.88$$

$$\hat{\sigma}^2 = \begin{pmatrix} 10.20 & -0.74 \\ -0.74 & 0.074 \end{pmatrix}$$

$Cov(\beta_0, \beta_1) \leftarrow -0.74$

$0.074 \rightarrow Cov(\beta_1, \beta_1) = \hat{\sigma}^2_{\beta_1}$

$$\begin{cases} Var(\beta_0) = 10.20 \\ Var(\beta_1) = 0.074 \end{cases}$$

Correlation $\rightarrow \rho(\beta_0, \beta_1)$

$$= \frac{Cov(\beta_0, \beta_1)}{\hat{\sigma}^2_{\beta_1}\hat{\sigma}^2_{\beta_0}} = \boxed{-0.98}$$