

(۱) اف  $f = \arg \min_f E[(f(x) - y)^2] = \arg \min_f \left[ \int_{-\infty}^{\infty} (f(x) - y)^2 f(y|x) dy \right]$

$$\arg \min_f \int_{-\infty}^{\infty} (f^2 - 2fy + y^2) f(y|x) dy = \arg \min_f \left[ f^2(x) \int_{-\infty}^{\infty} f(y|x) dy - 2f(x) \int_{-\infty}^{\infty} f(y|x) \cdot y dy + \int_{-\infty}^{\infty} y^2 f(y|x) dy \right]$$

$$= \arg \min_f \left[ f^2(x) - 2f(x) E(y|x) + \int_{-\infty}^{\infty} y^2 f(y|x) dy \right]$$

$$\frac{d}{df(x)} = 0 \quad 2f(x) - 2E(y|x) = 0 \rightarrow f(x) = E[y|x] \checkmark$$

(۱) ب  $P(y = y_k | X) \propto \exp(w_{k0} + \sum_{i=1}^d w_{ki} X_i)$  for  $k=1, \dots, K$

$$\Rightarrow P(y = y_k) = \frac{\exp(w_{k0} + \sum_{i=1}^d w_{ki} X_i)}{\sum_{k=1}^K \exp(w_{k0} + \sum_{i=1}^d w_{ki} X_i)}$$

در مدل های این چنینی که باید Bias که در اینجا  $w_{k0}$  است و یک سیگما از وزن ها و در هر جا تابعی دارد برای activation function که به دلیل وجود وزن ها باعث زیاده به شبکه های عصبی شود

(۱) ج برای طبقه بندی ب باید از آن  $\arg \max$  بگیریم و پس مشتق بگیریم

$$\arg \max \{ P(y = y_k | x) \} = \arg \max \left\{ \frac{\exp(w_{k0} + \sum_{i=1}^d w_{ki} X_i)}{\sum_{k=1}^K \exp(w_{k0} + \sum_{i=1}^d w_{ki} X_i)} \right\}$$

(۲) الف  $L_1$  regularization که کاربرد آن برای زمانی است که dimension زیاد باشد یا به نام لایم، قدر Feature ها را کم کند

$$L_1 \rightarrow \hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x_i \beta)^2 + \lambda \|\beta\|_1$$

$L_2$  Regularization برای dimension کم به جزای چون از کم feature ها استفاده می کند

$$L_2 \rightarrow \hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x_i \beta)^2 + \lambda \|\beta\|_2^2$$

این روش برای حل کردن بیشترین و به دست آوردن  $\beta$  استفاده می شود

$$\hat{\beta} = (A^T A + \lambda I)^{-1} A^T y$$



$$J(w) = \|Xw - y\|^2 + \lambda \|w\|^2 \Rightarrow \hat{w} = (x^T x + \lambda I)^{-1} x^T y \quad (2) \quad \text{ب}$$

برای min کردن  $J(w)$  از آن مشتق میگیریم

$$J(w) = (y - w^T x)^T (y - w^T x) + \lambda w^T w$$

$$\frac{dJ(w)}{dw} = -2x^T \cdot (y - w^T x) + 2\lambda w = 0 \Rightarrow 2\lambda w = 2x^T \cdot (y - w^T x)$$

$$\rightarrow \lambda w = x^T (y - w^T x) \rightarrow \lambda w = x^T y - x^T x w \rightarrow x^T y = x^T x w + \lambda w$$

$$\rightarrow (x^T x + \lambda I) w = x^T y \rightarrow w = (x^T x + \lambda I)^{-1} x^T y \quad \checkmark$$

(3) الف) از قبل فرمول  $\beta = (A^T A)^{-1} A^T y$  را داشته ایم پس در مسائل بعد از این فرمول استفاده

میکنیم چون در (الف) سبب حفظ در نظر نمیگیریم پس ماتریس  $A$  همی 1 خواهند بود

$$A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}_{10 \times 1} \rightarrow A^T A = [1 \dots 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 10 \rightarrow A^T y = [1 \dots 1] \begin{bmatrix} 44 \\ 20 \\ 30 \end{bmatrix} = 541$$

$$\beta_0 = (A^T A)^{-1} A^T y = \frac{1}{10} \times 541 = 54.1 \quad \checkmark$$

ب) در این بخش عرض از مبدأ داریم و سبب حفظ داریم پس ماتریس  $A$  برابر  $X$  می شود

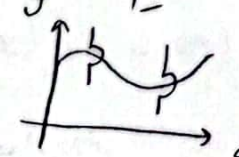
$$A = \begin{bmatrix} 14 \\ 27 \\ 10 \end{bmatrix} \rightarrow A^T A = [14 \ 27 \ 10] \begin{bmatrix} 14 \\ 27 \\ 10 \end{bmatrix} = 817 \quad \left. \begin{array}{l} A^T y = [14 \ 27 \ 10] \begin{bmatrix} 44 \\ 20 \\ 30 \end{bmatrix} = 12521 \end{array} \right\} \beta_1 = \frac{12521}{817} = 15.3 \quad \checkmark$$

ج)  $\epsilon$  همان نویز می باشد و در واقع در معادله دوم یعنی  $(y = \beta_0 + \beta_1 x + \epsilon)$  نویز را در نظر گرفته ایم

و در حالت اول  $(\hat{y} = b_0 + b_1 x)$  نویز در نظر نگرفته ایم

$$\epsilon \sim N(0, 0.2)$$

این نویز بر هر نقطه به تابع ما اضافه می شود یعنی



(د) اگر  $\alpha = 4$  را در معادله قرار دهیم خروجی 22 می شود اما این به این جهت نیست که سبب 22 خواهد شد به این دلیل که نویز می دهد و می تواند در دین ما حفظ شود

$$SSE \rightarrow \sum_{i=1}^{14} (y_i - \hat{y})^2 = 7 \quad \rightarrow \quad s^2 = \frac{\sum_{i=1}^{14} (y_i - \hat{y})^2}{n-1}$$

$$\rightarrow s^2 = \frac{7}{15} = 0.46 \quad \checkmark$$