عبادالدین رستمیان ۴۰۲۲۰۶۴۷۱

تمرین نظری سری اول

۱ الف) سافت‌مکس امتیازات خروجی مدل را به احتمالات تبدیل می‌کند که مجموع آن‌ها ۱ است. پس از آنجا که در کلاس‌بندی هدف پیش‌بینی احتمال هر کلاس است این تابع مناسب می‌باشد.

۱ ب) واریانس بالا یعنی مدل به نویزهای داده‌های آموزشی حساس است که باعث می‌شود مدل اورفیت کند. از روش‌های مرسوم برای کاهش واریانس می‌توان به Regularization اشاره کرد.

۱ ج) رگرسیون Ridge وزن‌ها را بین ویژگی‌ها توزیع می‌کند در حالی که رگرسیون Lasso ممکن است به برخی ویژگی‌ها وزن صفر اختصاص دهد و آن‌ها را حذف کند که در صورت همبستگی ویژگی‌ها به هم ایده‌آل نیست.

۱ د) رگولاریزیشن L2 یا همان رگرسیون Ridge با کاهش وزن‌های بزرگ از اورفیت جلوگیری می‌کند. اما اندکی به بایاس مدل می‌افزاید که نسبت به مدل رگولاریزنشده روی داده‌های آموزشی دقت کمتری بدست می‌آورد. ولی generalization مدل بالا رفته یعنی روی داده‌های دیده‌نشده دقت بهتری ارائه می‌دهد که مطلوب هم همین است.

۲ الف)

$$y = w^T x = w_1 x_1 + \dots + w_L x_L = x^T w$$

$$\rightarrow x y = x x^T w \xrightarrow[\text{j-th row}]{\text{only feature } j} x_j y_j = x_j x_j^T w_j$$

$$\rightarrow w_j = \frac{x_j y_j}{x_j x_j^T}$$

۲ ب) اگر دو ویژگی x_j و x_k مستقل باشند داریم:

$$x_j x_k^T = 0 \quad (j \neq k)$$

محاسبه وزن‌ها بر روی همه ویژگی‌ها:

$$w = (\underbrace{x x^T}_{\begin{bmatrix} x_1 x_1^T & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & x_L x_L^T \end{bmatrix}})^{-1} x y = \begin{bmatrix} \frac{1}{x_1 x_1^T} & 0 & \cdots & 0 \\ 0 & \frac{1}{x_2 x_2^T} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \frac{1}{x_L x_L^T} \end{bmatrix} x y$$

$$\Rightarrow w_j = \frac{x_j y}{x_j x_j^T}$$

$$y = W^T x + w_0 \longrightarrow W = (x x^T)^{-1} x^T (y - w_0) \longrightarrow w_j = \frac{x^T (y - w_0)}{x x^T}$$

$$X_{new} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1L} \\ \vdots & \vdots & x_{22} & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{N1} & \cdots & \cdots & x_{NL} \end{bmatrix}, \quad W_{new} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_L \end{bmatrix}$$

$$W_{new} = (X_{new} X_{new}^T)^{-1} X_{new} y$$

$$\mathbb{E}[y] = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$\mathbb{E}[\hat{y}] = \frac{1}{N} \sum_{i=1}^{N} (W^T x + w_0) = W^T \mathbb{E}[x] + w_0$$

$$\longrightarrow w_0 = \mathbb{E}[\hat{y}] - W^T \mathbb{E}[x]$$

$$x_2 \leq x_1 - 2 \rightarrow x_2 - x_1 \leq -2$$

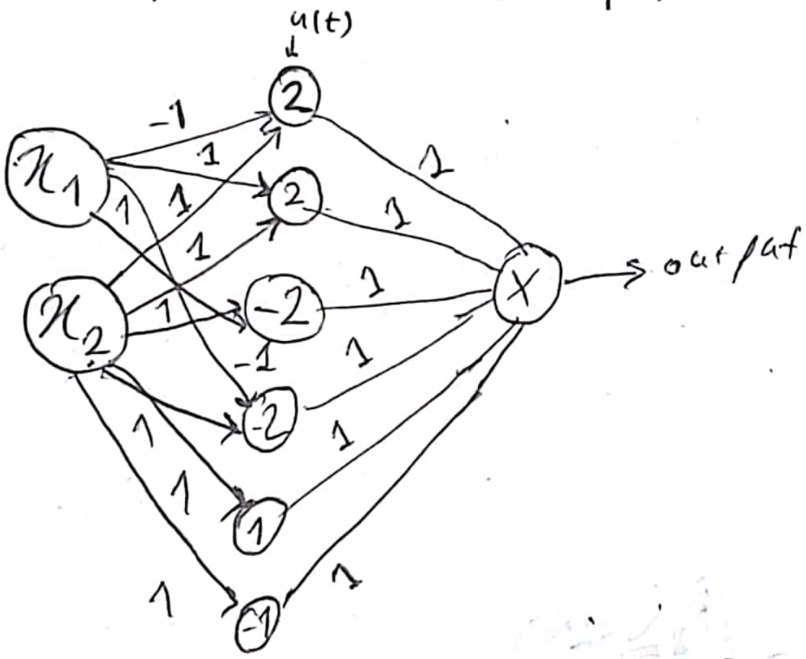
$$-1 \leq x_2 \leq 1$$

$$x_2 \leq -x_1 + 2 \rightarrow x_2 + x_1 \leq 2$$

$$\rightarrow -2 \leq x_1 + x_2 \leq 2$$

$$x_2 \geq -x_1 - 2 \rightarrow x_2 + x_1 \geq -2$$

$$x_2 \geq x_1 + 2 \rightarrow x_2 - x_1 \geq 2$$



$$i=k: \quad \frac{\partial \hat{y}_i}{\partial z_i} = \frac{\partial}{\partial z_i}\left(\frac{e^{z_i}}{\sum e^{z_j}}\right) = \frac{(e^{z_i} \cdot \sum e^{z_j}) - (e^{z_i} e^{z_i})}{(\sum e^{z_j})^2}$$ (الف ④

$$= \frac{e^{z_i}(\sum e^{z_j} - e^{z_i})}{(\sum e^{z_j})^2} = \hat{y}_i (1 - \hat{y}_i)$$

$$i \neq k: \quad \frac{\partial \hat{y}_i}{\partial z_k} = \frac{\partial}{\partial z_k}\left(\frac{e^{z_i}}{\sum e^{z_j}}\right) = \frac{-e^{z_i} e^{z_k}}{(\sum e^{z_j})^2} = -\hat{y}_i \hat{y}_k$$

(ب

$$\frac{\partial L}{\partial z_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial z_j}$$

$$\frac{\partial L}{\partial \hat{y}_j} = \frac{\partial}{\partial \hat{y}_j}\left(-\sum_{j=1}^{n} y_j \log(\hat{y}_j)\right) = -\frac{y_j}{\hat{y}_j}$$

$$\xRightarrow{(\text{الف})} \frac{\partial L}{\partial z_i} = -\frac{y_i}{\hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_i} - \sum_{j \neq i} \frac{y_j}{\hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial z_i}$$

$$= -y_i(1 - \hat{y}_i) + \sum_{j \neq i} y_j \hat{y}_i = \hat{y}_i \sum_{j=1}^{n} y_j - \hat{y}_i y_i$$

$$\sum_{j=1}^{n} y_i = 1 \longrightarrow \frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

5 الف)

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

$$Var(\hat{\beta}^{LS}) = \sigma^2 (X^T X)^{-1}$$

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

$$Var(\hat{\beta}^{Ridge}(\lambda)) = \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$$

λI باعث افزایش eigen value های ماتریس $X^T X$ می‌شود پس با inverse گرفتن مقدار eigen value ها کاهش می‌یابد پس:

$$(X^T X + \lambda I)^{-1} \leq (X^T X)^{-1}$$

$$\rightarrow \begin{cases} \lambda = 0 & Var(\hat{\beta}^{Ridge}) = Var(\hat{\beta}^{LS}) \\ \lambda > 0 & Var(\hat{\beta}^{Ridge}) < Var(\hat{\beta}^{LS}) \end{cases}$$

ب)

$$tr(Var[\hat{Y}(\lambda)]) = \sigma^2 \sum_{j=1}^{p} (D_x)^2_{jj} ((D_x)^2_{jj} + \lambda)^{-2}$$

$$X^T X = Q D_x Q^T$$

$$Var[\hat{Y}(\lambda)] = X(\underbrace{X^T X}_{Q D_x Q^T} + \lambda I)^{-1} \underbrace{Var(y)}_{\sigma^2} (X^T X + \lambda I)^{-1} X^T$$

$$= \sigma^2 Q (D_x + \lambda I)^{-1} D_x (D_x + \lambda I)^{-1} Q^T$$

$$\xrightarrow{tr} \sigma^2 \sum_{j=1}^{p} (D_x)^4_{jj} ((D_x)_{jj} [(D_x)_{jj} + \lambda)^{-2}$$