به نام خدا

پارسا پالیزیان - ۴۰۰۱۰۰۸۶۵

۱)

$$y_i = x_i^T \beta_0 + \varepsilon_i$$

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

$$R_{tr}(\hat{\beta}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \hat{\beta})^2 = \frac{1}{N} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$H = X (X^T X)^{-1} X^T$$

$$y - X\hat{\beta} = y - Hy$$

$$R_{tr}(\hat{\beta}) = \frac{1}{N} (y - Hy)^T (y - Hy)$$

$$H^T = X (X^T X)^{-1} X^T = H$$

$$H^2 = X \underbrace{(X^T X)^{-1} X^T X}_{I} (X^T X)^{-1} X^T = X (X^T X)^{-1} X^T = H$$

$$H^2 v = \lambda^2 v \xrightarrow{H^2 = H} Hv = \lambda^2 v \Rightarrow \lambda v = \lambda^2 v \Rightarrow$$

$$\lambda v (\lambda - 1) = 0 \Rightarrow \lambda = 0 \text{ یا } \lambda = 1$$

به تعداد رنک
$P = H$

$$tr(H) = rank(H) = p$$

به نام خدا

پارسا پالیزیان - ۴۰۰۱۰۰۸۶۵

۱)

$$E[R_{tr}(\hat{\beta})] = \frac{1}{N} E[tr((y-Hy)(y-Hy)^T)] \ *$$

$$= \frac{tr}{N}\left[E[(I-H)yy^T\underbrace{(I-H)^T}_{(I-H)}]\right] = \frac{tr}{N}\left[(I-H)E[yy^T](I-H)\right]$$

$$= \frac{tr}{N}\left[(I-H)[Var(y) + E[y]E[y]^T](I-H)\right]$$

$$= \frac{tr}{N}\left[(I-H)[\sigma^2 I + X\beta_0\beta_0^T X^T](I-H)\right]$$

$$= \frac{tr}{N}\left[[\sigma^2(I-H) + \underbrace{(I-H)X}_{0}\beta_0\beta_0^T X^T](I-H)\right]$$

$$= \frac{tr}{N}\left[\sigma^2(I-H)^2\right] = \frac{tr}{N}\left[\sigma^2(I-H)\right] = \frac{\sigma^2}{N}(N-p_0)$$

$$= \sigma^2\left(1-\frac{p}{N}\right) \leq \sigma^2 \overset{**}{\leq} E[R_{te}(\hat{\beta})]$$

$$* \ (y-Hy)^T(y-Hy) = tr[(y-Hy)^T(y-Hy)]$$

$$= tr[(y-Hy)(y-Hy)^T]$$

$**$

$$E[(\bar{y}-\bar{x}^T\hat{\beta})^2] = E[(\bar{x}^T\beta_0 + \bar{\varepsilon} - \bar{x}^T\hat{\beta})^2] = \qquad \geq \sigma^2$$

$$\underbrace{E[(X^T(\beta_0-\hat{\beta}))^2]}_{\geq 0} + \underbrace{E[\bar{\varepsilon}^2]}_{\sigma^2} + 2\underbrace{E[\bar{\varepsilon}]E[X^T(\beta_0-\hat{\beta})^2]}_{0}$$

$$E[R_{te}(\hat{\beta})] = \frac{1}{N}\sum_{i=1}^{\tilde{N}} E[(\bar{y}_i - \bar{x}^T\hat{\beta})^2] \geq \sigma^2$$

Scanned with CamScanner

به نام خدا

پارسا یالیزیان - ۴۰۰۱۰۰۸۶۵

۲) ۲ - ۲

$$Y = \beta_0 + \beta_1 X + \varepsilon \Rightarrow \hat{\beta}^{(lin)} = \underset{\beta}{\operatorname{argmin}} \|y - X_{linear}\beta\|_2^2$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \Rightarrow \hat{\beta}^{(Poly)} = \underset{\beta}{\operatorname{argmin}} \|y - X_{poly}\beta\|_2^2$$

$$\min_{\beta} \|y - X_{linear}\beta\|_2^2 = \|y - X_{linear}\hat{\beta}^{(lin)}\|_2^2$$

$$\beta^* = (\beta_0^{(lin)}, \beta_1^{(lin)}, 0, 0):$$

$$\|y - X_{poly}.\beta^*\|_2^2 = \|y - X_{linear}\hat{\beta}^{(lin)}\|_2^2$$

$$\|y - X_{poly}\hat{\beta}^{(Poly)}\|_2^2 \leqslant \|y - X_{poly}\beta^*\|_2^2 = \|y - X_{linear}\hat{\beta}^{(lin)}\|_2^2$$

$$\Rightarrow \min_{\beta} \|y - X_{poly}\beta\|_2^2 \leqslant \min_{\beta} \|y - X_{linear}\beta\|_2^2$$

به نام خدا

رسا پالیزبان - ۴۰۰۱۰۰۸۶۵

۲)

$$y = f(x) + \varepsilon$$

$$\hat{y}_1 = f_1(x)$$

$$\hat{y}_2 = f_2(x)$$

$$E_1 = \frac{1}{n} \sum_{i=1}^{n} \left(f(x) + \varepsilon - f_1(x) \right)^2$$

$$E_2 = \frac{1}{n} \sum_{i=1}^{n} \left(f(x) + \varepsilon - f_2(x) \right)^2$$

$$E = E\left[\left(f(x) + \varepsilon - \hat{f}(x)\right)^2 \right] = \iint p(x) \left(f(x) + \varepsilon - \hat{f}(x) \right)^2 dx\, d\varepsilon$$

در تعداد داده زیاد خطای هر دو تقریبا برابر است اما در تعداد داده کم تابع (مدل) پیچیده‌تر خطای کمتری روی داده‌های train دارد و overfit رخ می‌دهد.

برعکس در داده‌های test مدل درجه ۳ (که پیچیده‌تر است و overfit کرده) خطای بیشتری دارد.

$$n \to \infty : \iint \left(f(x) + \varepsilon - \hat{f}_1(x) \right)^2 dx\, d\varepsilon = \iint \left(f(x) + \varepsilon - \hat{f}_2(x) \right)^2 dx\, d\varepsilon = E$$

$$\left.\begin{array}{l} |E_1 - E| < \varepsilon \\ |E_2 - E| < \varepsilon \end{array}\right\} \Rightarrow |E_1 - E_2| = |E_1 - E - E_2 + E| \leq |E_1 - E| + |E_2 - E| < 2\varepsilon$$

بنام خدا

پارسا یالیزیان - ۴۰۰۱۰۰۸۶۵

۳) الف -

$$J_\lambda(w) = \frac{1}{2}(y - xw)^T(y - xw) + \lambda\|w\|_1$$

$$= \frac{1}{2}(y^Ty - y^TXw - w^TX^Ty + w^T\underbrace{X^TX}_{I.}w) + \lambda\|w\|_1$$

$$= \frac{1}{2}y^Ty + (-y^TXw + \frac{1}{2}w^Tw + \lambda\|w\|_1)$$

$$= \underbrace{\frac{1}{2}y^Ty}_{g(y)} + \sum_{i=1}^{d} \underbrace{\frac{1}{2}w_i^2 + \lambda|w_i| - y^TX_{:,i}w_i}_{f(X_{:,i}, y_i, w_i, \lambda)}$$

ب -

$$w_i \geq 0 \Rightarrow J_\lambda(w) = \frac{1}{2}y^Ty + \sum_{i=1}^{d}\frac{1}{2}w_i^2 + \lambda w_i - y^TX_{:,i}w_i$$

$$\frac{\partial J_\lambda(w)}{\partial w_i} = 0 \Rightarrow w_i + \lambda - y^TX_{:,i} = 0 \Rightarrow$$

$$w_i = y^TX_{:,i} - \lambda$$

$w_i \leq 0 \Rightarrow$ به طور مشابه $\Rightarrow w_i = y^TX_{:,i} + \lambda$

پ -

بنام خدا

پارسا یالنزیان - ۴۰۰۱۰۰۸۶۵

۳) ت - در صورتی w_i ها صفر می‌شود که در هر دو رابطه بالا صدق کند یعنی:

$$y^T X_{:,i} - \lambda = y^T X_{:,i} + \lambda \Rightarrow 2\lambda = 0 \Rightarrow \lambda = 0$$

$$\underbrace{y^T X_{:,i} - \lambda > 0}_{w_i > 0} \Rightarrow y^T X_{:,i} > \lambda \text{ و } \underbrace{y^T X_{:,i} < -\lambda}_{w_i < 0} \Rightarrow |y^T X_{:,i}| < \lambda \Rightarrow w_i = 0$$

ث -

$$J_\lambda(w) = \frac{1}{2}\|y - Xw\|_2^2 + 0.5\lambda\|w\|_2^2$$

$$= \frac{1}{2}(y - Xw)^T(y - Xw) + \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}y^T y - y^T X w + \frac{1}{2}w^T \underbrace{X^T X}_{I} w + \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}y^T y + \sum_{i=1}^{d} \frac{1}{2}(\lambda+1)w_i^2 - y^T X_{:,i} w_i$$

$$\frac{\partial J_\lambda(w)}{\partial w_i} = 0 \Rightarrow (\lambda+1)w_i - y^T X_{:,i} = 0 \Rightarrow$$

$$w_i = \frac{y^T X_{:,i}}{\lambda + 1}$$

زمانی که $y^T X_{:,i}$ صفر یا $\lambda \to \infty$، w_i ها صفر می‌شوند. $y^T X_{:,i}$ دست ما نیست و λ نیز در عمل ∞ نیست بنابراین در این روش احتمال اینکه w_i ها صفر (sparse) شوند، کمتر است.

بنام خدا

پارسا پالیزیان - ۴۰۰۱۰۰۸۷۵

۴) الف -

$$\varepsilon_i < 0 \Rightarrow y^{(i)}(w^T x^{(i)} + b) \geq 1 - \varepsilon_i$$ اگر

معادل است با

$$y^{(i)}(w^T x^{(i)} + b) \geq 1$$

در نتیجه چون می‌خواهیم $\min \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \varepsilon_i^2$ را حل کنیم این عبارت سعی می‌کند ε_i ها را صفر کند و برای ε_i های منفی آن ها را دقیقاً صفر می‌کند و همان طور که نشان دادیم با constraint های مسئله در تناقض نمی‌شود بنابراین ε_i را بگذاریم هم خود به خود $\varepsilon_i \geq 0$ می‌شود.

ب -

$$L(w, \alpha, \varepsilon, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \varepsilon_i^2$$

$$- \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 + \varepsilon_i \right]$$

ج -

شرایط KKT :

$$\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\nabla_\varepsilon L = C\varepsilon - \alpha = 0 \Rightarrow C\varepsilon_i = \alpha_i$$

$$\alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 + \varepsilon_i \right] = 0$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \varepsilon_i$$

$$\alpha_i \geq 0$$

Scanned with CamScanner

به نام خدا

پارسا پالیزیان - ۴۰۰۱۰۰۸۶۵

۴) د-

$$L = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)} +$$

$$\frac{1}{2}\sum_{i=1}^{m}\alpha_i\varepsilon_i - \sum_{i=1}^{m}\alpha_i\varepsilon_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y^{(i)} b$$

$$- \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)} =$$

$$\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)} - \frac{1}{2}\sum_{i=1}^{m}\frac{\alpha_i^2}{C}$$

مسئله dual به صورت زیر تعریف می‌شود:

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)} - \frac{1}{2}\sum_{i=1}^{m}\frac{\alpha_i^2}{C}$$

$$\text{s.t.} \quad \forall\, i = 1, 2, \dots, m: \begin{cases} \alpha_i \geq 0 \\ \sum \alpha_i y^{(i)} = 0 \end{cases}$$

بنام خدا

پارسا پالیزیان - ۴۰۰۱۰۰۸۶۵

۵) الف -

$$\nabla_{\beta} J = 0 \Rightarrow -A^T 2W(Y - A\beta) = 0 \Rightarrow$$

$$A^T W Y - A^T W A \beta = 0 \Rightarrow \beta = (A^T W A)^{-1} A^T W Y$$

ب -

$$A^T W A \in \mathbb{R}^{2\times 2}$$

$$\operatorname{rank}(A^T W A) = \begin{cases} 2 & \text{اگر همهٔ } x \text{ ها باهم برابر نباشند} \\ 1 & \text{,, ,, ,, باشند} \end{cases}$$

$\Rightarrow$ جواب منحصر بفرد داشت زیرا $A^T W A$ وارون دارد.

$\Rightarrow$,, ,, ,, نیست ,, ,, ندارد.

ج -

$$\beta(t+1) = \beta(t) - \alpha \nabla_{\beta} J(t) = \beta(t) + \alpha\left(A^T 2W(Y - A\beta(t))\right)$$

α: step-size

بنام خدا

پارسا یالیزیان - ۴۰۰۱۰۰۸۶۵

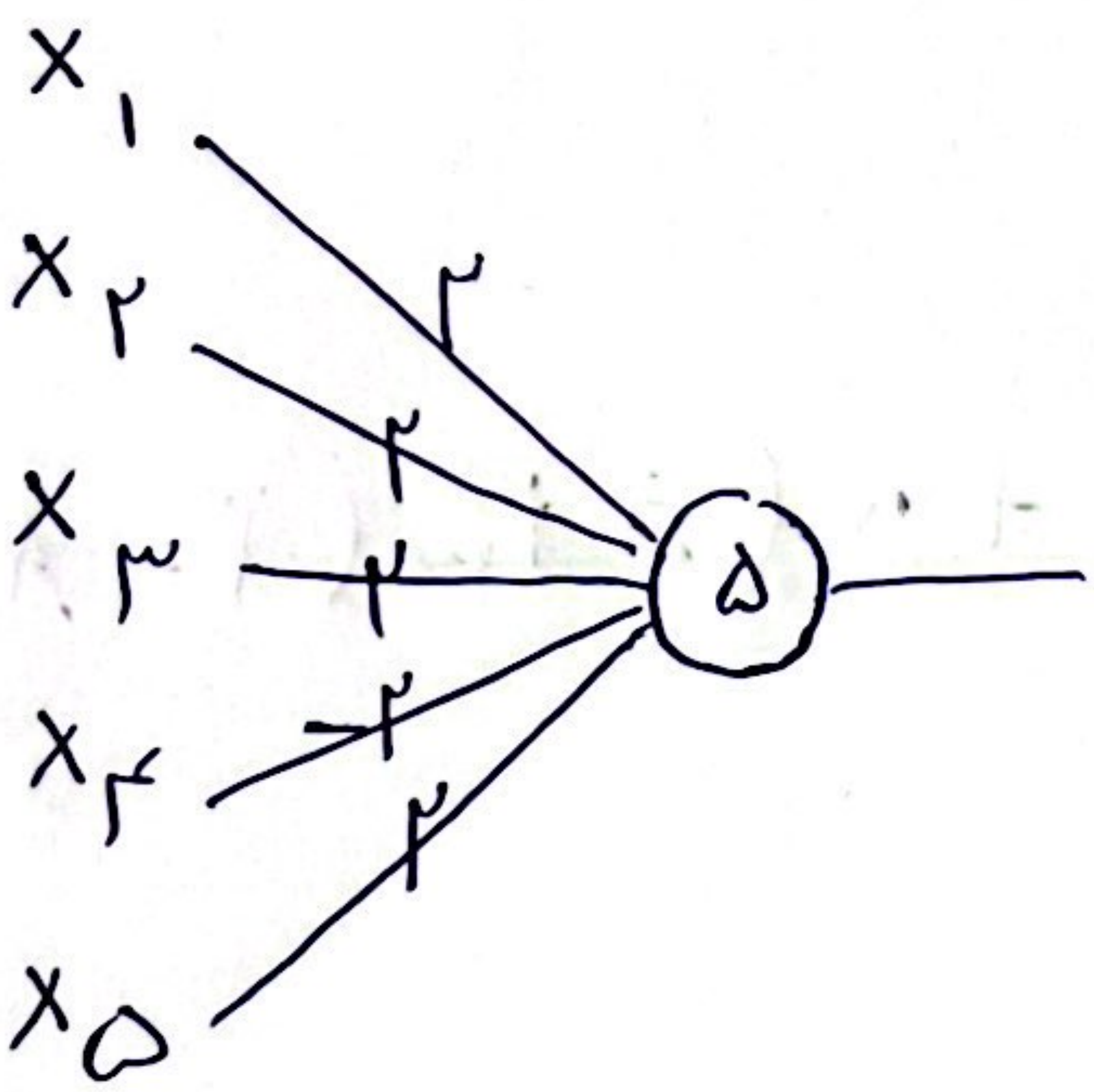
۶) الف- ۵ نورون ورودی (برای ۵ بیت) و ۱ نورون خروجی (اگر بخواهیم خروجی را باینری کد کنیم در غیر این صورت می‌توانیم ۲ نورون خروجی داشته باشیم و خروجی را به صورت one-hot کد کنیم)

ب- بله می‌توان. زیرا می‌توان مسئله را به صورت روبه رو مدل کرد و شبکه یک لایه MLP آن در قسمت بعدی رسم شده است.

۱۱۱۰۱ —(فاصله همینگ = ۰)→ ۱۱۱۰۱

۱۱۱۰۱ —(فاصله همینگ = ۱)→ {۰۱۱۰۱، ۱۰۱۰۱، ۱۱۰۰۱، ۱۱۱۱۱، ۱۱۱۰۰}

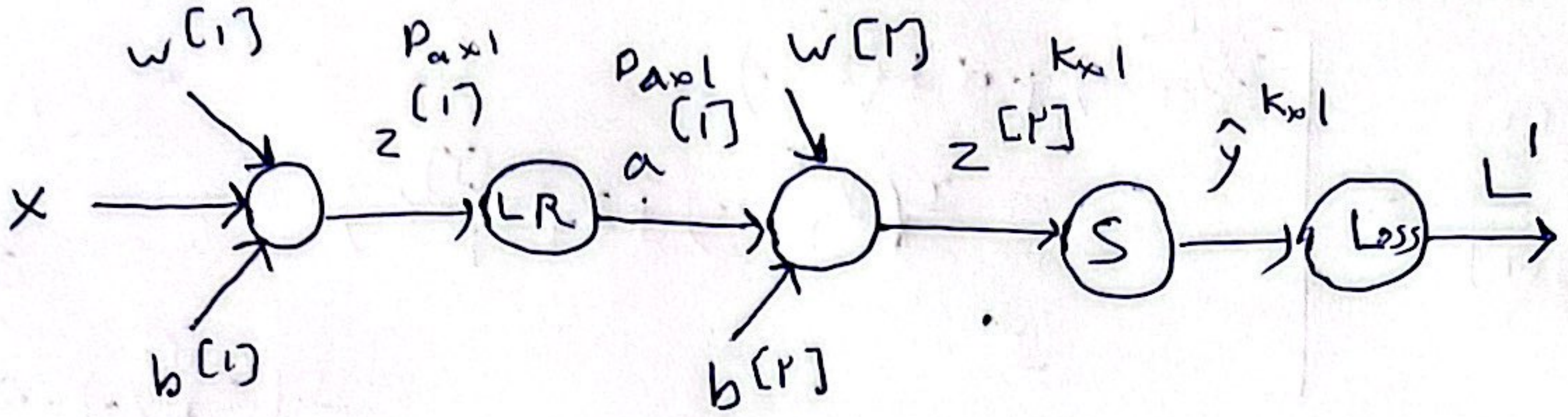
ج-



بنام خدا

پارسا یالنزیان - ۴۰۰۱۰۰۸۶۵

۷) الف -



$$W^{[2]} \in \mathbb{R}^{K \times p_a} \quad , \quad b^{[2]} \in \mathbb{R}^{K \times 1}$$

$$z^{[1]} \in \mathbb{R}^{p_a \times m}$$

ب -

$$\frac{\partial \hat{y}_K}{\partial z_K^{[2]}} = \frac{-e^{z_K^{[2]}} e^{z_K^{[2]}} + e^{z_K^{[2]}} \sum_{j=1}^{K} e^{z_j^{[2]}}}{\left(\sum_{j=1}^{K} e^{z_j^{[2]}}\right)^2} = \hat{y}_K - \hat{y}_K^2$$

$$\hat{y}_K = \frac{e^{z_K^{[2]}}}{\sum_{j=1}^{K} e^{z_j^{[2]}}}$$

ج -

$$\frac{\partial \hat{y}_K}{\partial z_i^{[2]}} = \frac{-e^{z_i^{[2]}} e^{z_K^{[2]}}}{\left(\sum_{j=1}^{K} e^{z_j^{[2]}}\right)^2} = -\hat{y}_i \hat{y}_K$$

د -

$$\frac{\partial L}{\partial z_i^{[2]}} = \frac{\partial \hat{y}_i}{\partial z_i^{[2]}} \frac{\partial L}{\partial \hat{y}_i} = \hat{y}_i (1 - \hat{y}_i) \frac{-y_i}{\hat{y}_i} = \hat{y}_i - 1$$

$$\frac{\partial L}{\partial z_i^{[2]}} = \frac{\partial \hat{y}_K}{\partial z_i^{[2]}} \frac{\partial L}{\partial \hat{y}_K} = -\hat{y}_i \hat{y}_K \frac{-y_K}{\hat{y}_K} = \hat{y}_i$$

ه -

$$\delta_1 = \frac{\partial z^{[2]}}{\partial a^{[1]}} = W^{[2]T}$$

به نام خدا

پارسا پالیزیان - ۴۰۰۱۰۰۸۲۵

۷) و-

$$\delta_2 = \frac{\partial a^{[1]}}{\partial z^{[1]}} = \begin{bmatrix} I(z_1^{[1]}>0)+\alpha I(z_1^{[1]}<0) & & 0 \\ & I(z_2^{[1]}>0)+\alpha I(z_2^{[1]}<0) & \\ 0 & & \ddots \\ & & & I(z_n^{[1]}>0)+\alpha I(z_n^{[1]}<0) \end{bmatrix}$$

ز-

$$\frac{\partial L}{\partial w^{[1]}} = \underbrace{\frac{\partial z^{[1]}}{\partial w^{[1]}}}_{x^T} \underbrace{\frac{\partial a^{[1]}}{\partial z^{[1]}}}_{\delta_2} \underbrace{\frac{\partial z^{[2]}}{\partial a^{[1]}}}_{\delta_1} \underbrace{\frac{\partial \hat{y}}{\partial z^{[2]}} \frac{\partial L}{\partial \hat{y}}}_{\delta_0}$$

$$= \delta_2 \delta_1 \delta_0 x^T$$

ح- مشکل اولیه softmax در عمل این است که چون ممکن است $z_j^{[2]}$ ها خیلی بزرگ باشند در نتیجه $e^{z_j^{[2]}}$ خیلی خیلی بزرگ می‌شود و overflow رخ می‌دهد در نتیجه یک مقدار DC که برابر بزرگ‌ترین مقدار بین $z_j^{[2]}$ ها است از همه‌ی مقادیر کم می‌شود تا $\max_j e^{z_j^{[2]}} = 1$ شود و overflow اتفاق نیفتد.

دقت کنید:

$$\hat{y}_{i\,\text{جدید}} = \frac{e^{z_i^{[2]}-m}}{\sum_{j=1}^{K}\left(e^{z_j^{[2]}-m}\right)} = \frac{e^{-m} e^{z_i^{[2]}}}{e^{-m}\sum_{j=1}^{K} e^{z_j^{[2]}}}$$

$$= \frac{e^{z_i^{[2]}}}{\sum_{j=1}^{K} e^{z_j^{[2]}}} = \hat{y}_{i\,\text{قبلی}}$$

در نتیجه ما ساختار مدل را تغییر ندادیم و می‌توانیم از $\hat{y}_i$ جدید استفاده کنیم.