به نام خدا

علی قاسم زاده 401106339

۱-

الف) softmax بصورت smooth، بزرگترین احتمال را برجسته می‌کند، همچنین نسبت به تابع max که فاقد مشتق است، بهتر است زیرا مشتق پذیر است، softmax مقادیر منفی را هم هندل می‌کند زیرا تابع نمایی است و e بتوان عدد منفی باز هم مثبت است.

ب) واریانس بالا به صورت کلی یعنی مدل پیچیده، وقتی که دیتامون نسبت به پیچیدگی مدل کم باشه هم این واریانس بالا باعث overfitting می‌شود. در واقع وقتی فضای فرضیه بزرگی داشته باشیم یعنی مدل پیچیده که واریانس زیاد است. یه روش برای کاهش واریانس در مدل، کم کردن پیچیدگی مدل است.

پ) زیرا رگرسیون Lasso می‌خواهد تعدادی از feature ها را حذف کند و با این کار sparsity را زیاد کند، در صورتی که اگر تمام ویژگی‌ها تا حد خوبی با خروجی مرتبط باشند، باعث خطای بزرگی می‌شود ولی Ridge، ویژگی‌ها را حذف نمی‌کند (منظور از حذف ویژگی، در نظر گرفتن وزن صفر برای آن است.) بلکه متناسب با اهمیت ویژگی‌ها به آنها وزن کمتر یا بیشتر می‌دهد و وقتی که همه‌ی ویژگی‌ها تا حد خوبی با خروجی مرتبط باشند وزن های مخالف صفر به آنها می‌دهد و در نتیجه هیچکدام از ویژگی ها ایگنور نمی‌شوند و مدل بهتر عمل می‌کند.

ت) L2 Regularization باعث می‌شود ضرایب مون تا حدی کوچک شوند ← باعث کوچک شدن فضای فرضیه و کم شدن Variance می‌شود (چون وزن های بزرگ حالت های خاص حذف می‌شوند، فضای فرضیه کوچک می‌شود) از طرفی چون فضای فرضیه مون کوچک شده، بایاس مون زیاد میشه ← در کل باعث کاهش واریانس و افزایش بایاس می‌شود

ادامه ی ب)

در واقع واریانس بالا نشانه ی اورفیت شدن و وابسته شدن زیاد مدل به دیتا ها است و مدل تعمیم پذیری خوبی نخواهد داشت

۲- تابع loss مان را mse در نظر می‌گیریم، داریم که:

$$\boxed{y_k = w^T x_k = x_k^T w \Rightarrow y = X^T w}$$

$$Loss = \frac{1}{n} \| y - X^T w \|_2^2$$

حالا برای w_j داریم که:

$$SSE(w_j) = \| y - w_j X_j^T \|_2^2 = (y - w_j X_j^T)^T (y - w_j X_j^T)$$

$$\rightarrow \frac{\partial SSE(w_j)}{\partial w_j} = -2 X_j (y - w_j X_j^T) = 0 \rightarrow X_j y - w_j X_j X_j^T = 0$$

$$\rightarrow \boxed{w_j = \frac{X_j y}{X_j X_j^T}}$$

ب) برای بهینه‌سازی روی کل $Loss = \frac{1}{n} \| y - X^T w \|_2^2$ باید این را کمینه کنیم

$$\nabla_w Loss = \frac{1}{n} \nabla_w \left((y - X^T w)^T (y - X^T w) \right) = 0 \rightarrow X (y - X^T w) = 0$$

$$\rightarrow X y - X X^T w = 0 \rightarrow X y = (X X^T) w$$

طبق گفته سوال و کوئرا: سطرهای X بر هم عمود اند ← داریم که:

$$X X^T = \begin{bmatrix} X_1 X_1^T & & \overset{\text{صفر}}{0} \\ & \ddots & \\ \underset{\text{صفر}}{0} & & X_n X_n^T \end{bmatrix} \rightarrow (X X^T)^{-1} = \begin{bmatrix} \frac{1}{X_1 X_1^T} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{X_n X_n^T} \end{bmatrix}$$

$$\rightarrow w = (X X^T)^{-1} X y = \begin{bmatrix} \frac{1}{X_1 X_1^T} & & \\ & \ddots & \\ & & \frac{1}{X_n X_n^T} \end{bmatrix} \begin{bmatrix} X_1 y \\ \vdots \\ X_n y \end{bmatrix} = \begin{bmatrix} \frac{X_1 y}{X_1 X_1^T} \\ \vdots \\ \frac{X_n y}{X_n X_n^T} \end{bmatrix} \rightarrow w_j = \frac{X_j y}{X_j X_j^T}$$

← w_j های حاصل از بهینه سازی روی تک متغیر، با w_j های حاصل از بهینه سازی روی کل برابر است.

ب) در این مورد باید در نظر بگیریم:

$$Loss(w_j, w_0) = SSE(w_j, w_0) = \| Y - w_j X_j^T - \tilde{W}_0 \|_2^2$$

بردار هم سایز بالا با تمام عناصر w_0

$$\rightarrow \frac{\partial Loss}{\partial w_j} = \frac{\partial}{\partial w_j}\left((Y - w_j X_j^T - \tilde{w}_0)^T (Y - w_j X_j^T - \tilde{w}_0)\right) =$$

$$X_j (Y - w_j X_j^T - \tilde{w}_0) = 0 \rightarrow X_j Y - w_j X_j X_j^T - X_j \tilde{W}_0 = 0$$

$$\rightarrow \boxed{w_j X_j X_j^T = X_j Y - X_j \tilde{w}_0} \rightarrow \boxed{w_j = \frac{X_j Y - X_j \tilde{w}_0}{X_j X_j^T}}$$

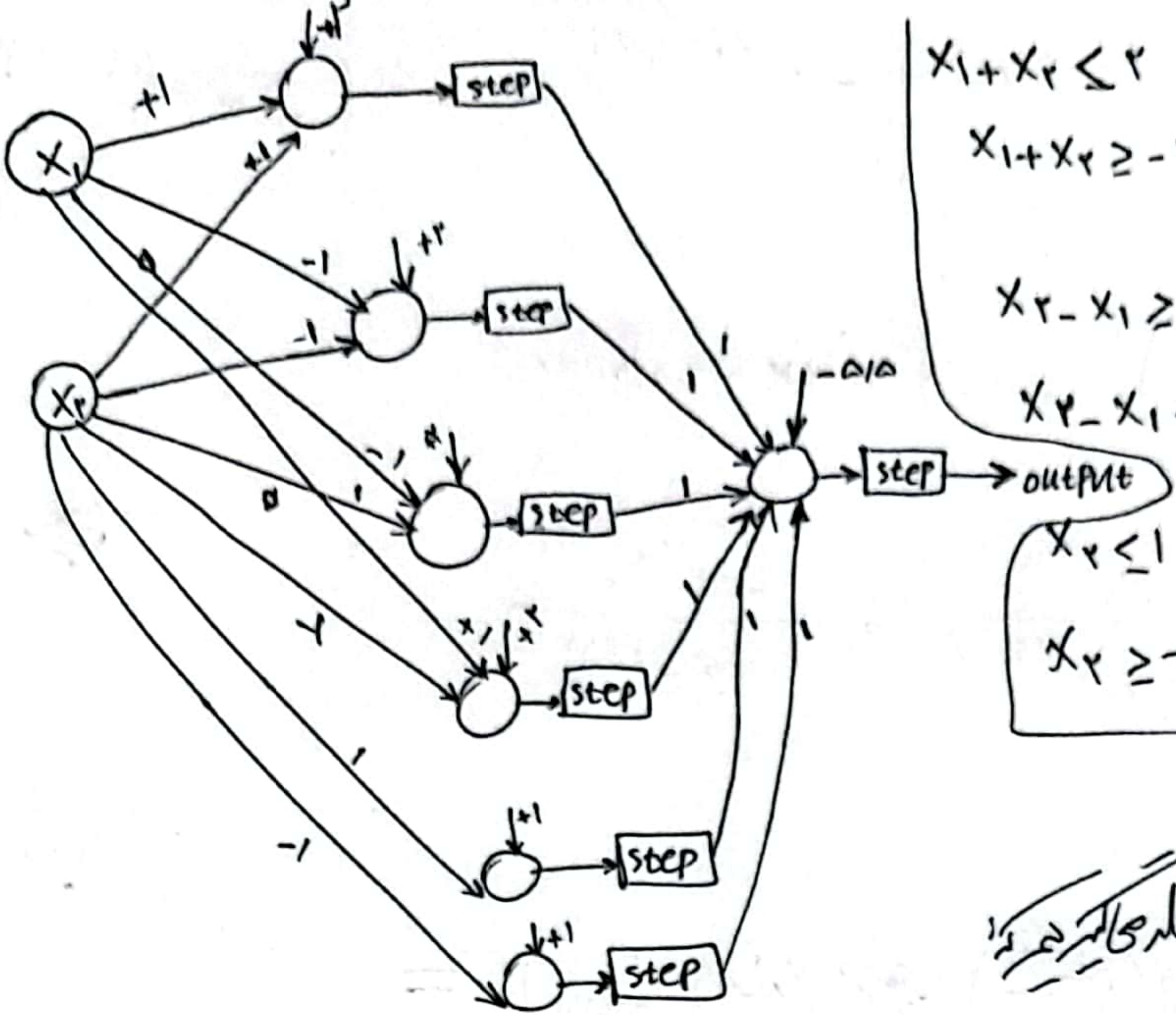
$$\frac{\partial Loss}{\partial w_0} = \frac{\partial}{\partial w_0}\left((Y - w_j X_j^T - \tilde{w}_0)^T (Y - w_j X_j^T - \tilde{w}_0)\right) = 0 \rightarrow$$

$$[1 \cdots 1](Y - w_j X_j^T - \tilde{w}_0) = 0 \rightarrow 1^T Y - w_j 1^T X_j^T - n w_0 = 0$$

$$\rightarrow n w_0 = 1^T Y - w_j 1^T X_j^T \rightarrow \boxed{w_0 = \frac{1^T Y - w_j 1^T X_j^T}{n}}$$

$$\rightarrow \boxed{w_0 = \frac{\sum y_i - w_j \sum x_{ji}}{n}}$$

-۳



$x_1 + x_2 \leq 2$

$x_1 + x_2 \geq -2$

$x_2 - x_1 \geq -2$

$x_2 - x_1 \leq 2$

$x_2 \leq 1$

$x_2 \geq -1$

step function مان را اینگونه در نظر می‌گیریم

$$f(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

۴- الف)

$$i=k: \frac{\partial}{\partial z_i}\left(\frac{e^{z_i}}{\sum e^{z_j}}\right) = \frac{e^{z_i}\left(\sum e^{z_j}\right) - e^{z_i}e^{z_i}}{\left(\sum e^{z_j}\right)^2} =$$

$$\frac{e^{z_i}}{\sum e^{z_i}} - \left(\frac{e^{z_i}}{\sum e^{z_j}}\right)^2 = \text{softmax} - \text{softmax}^2$$

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

ب)

$$\frac{\partial L}{\partial z_i} = -y_i \frac{\partial}{\partial z_i} \log\left(\frac{e^{z_i}}{\sum e^{z_j}}\right) - \frac{\partial}{\partial z_i}\sum_{k\neq i} y_k \log\left(\frac{e^{z_k}}{\sum e^{z_j}}\right) =$$

$$-y_i\left(\frac{\partial}{\partial z_i}\left(z_i - \log\sum e^{z_j}\right)\right) - \sum_{k\neq i} y_k \frac{\partial}{\partial z_i}\left(z_k - \log\sum e^{z_j}\right) =$$

$$-y_i\left(1 - \frac{e^{z_i}}{\sum e^{z_j}}\right) - \sum_{k\neq i} y_k\left(-\frac{e^{z_i}}{\sum e^{z_j}}\right) = -y_i\left(1 - \frac{e^{z_i}}{\sum e^{z_j}}\right) + \sum_{k\neq i} y_k\left(\frac{e^{z_i}}{\sum e^{z_j}}\right)$$

$$= -y_i + y_i\frac{e^{z_i}}{\sum e^{z_j}} + \sum_{k\neq i} y_k \frac{e^{z_i}}{\sum e^{z_j}} = -y_i + \underbrace{\sum_k y_k \frac{e^{z_i}}{\sum e^{z_j}}}_{\frac{e^{z_i}}{\sum e^{z_j}}}$$

$$= -y_i + \frac{e^{z_i}}{\sum e^{z_j}}$$

Scanned with CamScanner

۵- الف) مدل رگرسیون با خطا ε را در نظر بگیرید،

$$Y = Xw + \varepsilon$$

و $E(\varepsilon|X)=0$, $Var(\varepsilon|X)=\sigma^2 I$.

در این صورت می‌دانیم که جواب مسئله در حالت خطی و ridge عبارت است از:

$$\hat{w}_{LS} = (X^TX)^{-1}X^Ty$$

$$\hat{w}_{ridge} = (X^TX+\lambda I)^{-1}X^Ty$$

حالا داریم بایاس‌ها را حساب کنیم،

$$E(\hat{w}_{LS}|X) = E\left((X^TX)^{-1}X^T(Xw+\varepsilon)\,|X\right) = (X^TX)^{-1}X^TXw$$

$$E(\hat{w}_{ridge}|X) = E\left((X^TX+\lambda I)^{-1}X^T(Xw+\varepsilon)\,|X\right) = (X^TX+\lambda I)^{-1}X^TXw$$

ماتریس‌های کوواریانس‌ها نیز به صورت زیر است،

$$Var(\hat{w}_{LS}|X) = E\left((\hat{w}_{LS}-E(\hat{w}_{LS}|X))^2\,|X\right) = E\left(((X^TX)^{-1}X^T\overset{(Xw+\varepsilon)}{y} - (X^TX)^{-1}X^TXw)^2\,|X\right)$$

$$= (X^TX)^{-1}X^T\underbrace{E(\varepsilon^2|X)}_{\sigma^2 I}X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

$$Var(\hat{w}_{ridge}|X) = E\left((\hat{w}_{ridg} - E(\hat{w}_{ridge}|X))^2\,|X\right) =$$

$$E\left(((X^TX+\lambda I)^{-1}X^T\underset{Xw+\varepsilon}{y} - (X^TX+\lambda I)^{-1}X^TXw)^2\,|X\right) = \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}$$

حال قرار دهید

$$B = X^TX(X^TX+\lambda I)^{-1}$$

$$\Rightarrow var(\hat{w}_{ridge}|X) = \sigma^2 B^T (X^TX)^{-1} B$$

حال باید ثابت کنیم $var(\hat{w}_{ls}|X) - var(\hat{w}_{ridge}|X)$ یک ماتریس مثبت معین است.

← اعمال روی فضای مثبت و در نتیجه var_{ridge} کمتر از var_{ls} است.

$$var(\hat{w}_{ls}|X) - var(\hat{w}_{ridge}|X) = \sigma^2(X^TX)^{-1} - \sigma^2 B^T (X^TX)^{-1}B$$

$$= \sigma^2\left(B^T(B^T)^{-1}(X^TX)^{-1}B^{-1}B - B^T(X^TX)^{-1}B\right) =$$

$$\sigma^2 B^T\left((B^T)^{-1}(X^TX)^{-1}B^{-1} - (X^TX)^{-1}\right)B =$$

$$= \sigma^2 B^T\left(2\lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3}\right)B =$$

$$\sigma^2 (X^TX+\lambda I)^{-1}X^TX\left(2\lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3}\right)X^TX(X^TX+\lambda I)^{-1} =$$

$$\sigma^2 (X^TX+\lambda I)^{-1}\left(2\lambda I + \lambda^2(X^TX)^{-1}\right)(X^TX+\lambda I)^{-1}$$

حال بردار ناصفر v را از دو طرف اگر در این ماتریس ضرب کنیم، داریم

$$\sigma^2 \underbrace{v^T(X^TX+\lambda I)^{-1}}_{Z^T}\left(2\lambda I + \lambda^2(X^TX)^{-1}\right)\underbrace{(X^TX+\lambda I)^{-1}v}_{Z} =$$

$$\sigma^2 Z^T\left(2\lambda I + \lambda^2(X^TX)^{-1}\right)Z = 2\sigma^2 Z^TZ + \sigma^2\lambda^2 Z^T(X^TX)^{-1}Z$$

که X^TX یک ماتریس مثبت معین است $\Leftarrow v^TX^TXv>0$

به نرم 2

$$2\sigma^2 Z^TZ+\sigma^2\lambda^2 Z^T(X^TX)^{-1}Z \text{ is positive definite}$$

$\Leftarrow$

$$Var_{LS} > Var_{ridge}$$

ب)

$$var(\hat{Y}) = var(Xw)$$

$$tr\{Var[\hat{Y}(\lambda)]\} = \sigma^2\, tr\{X(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}X^T\}$$

$$X=U\overset{D}{D}V^T \longrightarrow X^TX = VD^2V^T \longrightarrow X^TX = VD^2V^T+\lambda I = VD^2V^T+\lambda VV^T$$

$$= V(D^2+\lambda I)V^T \longrightarrow (X^TX+\lambda I)^{-1} = V(D^2+\lambda I)^{-1}V^T$$

$$tr\{var[\hat{Y}(\lambda)]\} = \sigma^2\, tr\{\overset{UDV^T}{V(D^2+\lambda I)^{-1}}\underbrace{V^TV}_{I}D^2\underbrace{V^TV}_{I}(D^2+\lambda I)^{-1}V^T\overset{VDU^T}{}\} =$$

$$\sigma^2\, tr\{\overset{UD}{}(D^2+\lambda I)^{-1}D^2(D^2+\lambda I)^{-1}\overset{D}{U^T}\} = \sigma^2\, tr\{\underset{D}{}(D^2+\lambda I)^{-1}D^2(D^2+\lambda I)^{-1}\underset{D}{}\}$$

$$= \sigma^2\, tr\{(D^2+\lambda I)^{-2}D^4\} = \sigma^2\sum D_{jj}^4(D_{jj}^2+\lambda)^{-2}$$