علیرضا دهقانی پورفراشاه ۹۸۱۰۱۵۵۵

سؤال ۱:

Part1: گزینه c. بزرگ بودن ضریب به معنای مهم بودن و تأثیر زیاد آن روی مدل است ولی ممکن است با یک ویژگی دیگر وابستگی بالایی داشته باشد و مثلاً آن ویژگی بتواند این را خنثی کند و تأثیر آن را از مدل از بین ببرد بنابراین درمورد اهمیت در مدل و امکان نگه‌داری یا حذف آن نمی‌توان نظر قطعی داد.

Part2:

a) درست، مثلاً زمانی که مدل ما ساده است و underfit شده است با افزایش داده نمی‌توان این مشکل را برطرف نمود.

b) نادرست، مثلاً ممکن است مدل پیچیده باشد و ما به شدت overfit شویم و خطای train کم شود ولی خطای تست به شدت زیاد شود.

c) نادرست، خطای train کم می‌شود ولی اگر داده برای train زیاد باشد و حتی اگر خود مدل هم پیچیده باشد خطای test نیز کم خواهد شد.

d) نادرست، ممکن است مدل پیچیده‌تر باشد و نیاز شود از یک چندجمله‌ای با درجه‌ی بالاتر استفاده شود.

علیرضا دهقان پور فراشاه ۹۸۱۰۱۵۵۵

سوال ۲: $L(w) = \frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)^T}w - y^{(i)}\right)^2 = (Xw-Y)^T(Xw-Y)$

a)

$$\Rightarrow L(w) = w^TX^TXw - w^TX^TY - Y^TXw - Y^TY$$

$$\frac{\partial L}{\partial w} = 2X^TXw - X^TY - X^TY = 0 \Rightarrow 2X^TXw = 2X^TY \Rightarrow$$

$$\Rightarrow X^TXw = X^TY \Rightarrow \boxed{w^* = (X^TX)^{-1}X^TY}$$

b)

$$L(w) = (Xw-Y)^T(Xw-Y) + \lambda w^Tw$$

$$\frac{\partial L}{\partial w} = 2X^TXw - 2X^TY + 2\lambda w = 0 \Rightarrow (X^TX+\lambda I)w = X^TY$$

$$\Rightarrow w^* = (X^TX+\lambda I)^{-1}X^TY$$

c) ① ابتدا رفت را اثبات می‌کنیم یعنی اگر $\Sigma X = XF$ و F وارون پذیر:

$$\Sigma X = XF \xRightarrow{T} X^T\Sigma^T = F^TX^T \xRightarrow{\Sigma^T=\Sigma} X^T\Sigma = F^TX^T \xRightarrow{\times\Sigma^{-1}}$$

$$X^T = F^TX^T\Sigma^{-1} \xRightarrow{F^{-T}\times} F^{-T}X^T = X^T\Sigma^{-1} \quad *$$

در w^*_{new} به جای $X^T\Sigma^{-1}$ عبارت $F^{-T}X^T$ می‌گذاریم:

$$w^*_{new} = (F^{-T}X^TX)^{-1}F^{-T}X^TY = (X^TX)^{-1}F^T\times F^{-T}X^TY = \underbrace{(X^TX)^{-1}XY}_{w^*_{opt}}$$

② حال برگشت را اثبات می‌کنیم:

فرض: $(X^TX)^{-1}X^TY = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$

$$\Rightarrow \left[(X^TX)^{-1}X^TY - (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right]Y = 0$$

چون Y لیبل داده‌ها است و برای هر دیتاست این تساوی برقرار است Y نمی‌تواند در فضای پوچ (null space) ماتریس سمت چپ باشد بنابراین ماتریس سمت چپ صفر است یعنی:

$$(X^TX)^{-1}X^T = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$$

ادامه اثبات صفحه بعد:

ادامه سؤال ۲:

$$(X^TX)^{-1}X^T = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$$

$$\underbrace{(X^T\Sigma^{-1}X)\,(X^TX)^{-1}}_{A}\,\underbrace{X^T} = X^T\Sigma^{-1}$$

ماتریس A حاصل ضرب ۲ ماتریس وارون پذیر است پس وارون پذیر است.

$$AX^T = X^T\Sigma^{-1} \xRightarrow{\times\Sigma} AX^T\Sigma = X^T \xrightarrow{\text{ترنسپوز}} \Sigma^T X A^T = X \Rightarrow$$

$$\xRightarrow{\Sigma\ \text{متقارن}} \Sigma X A^T = X \underset{\text{A وارون پذیر}}{\Longrightarrow} \Sigma X = X\underbrace{A^{-T}}_{F} = XF$$

پس ثابت کردیم F وجود دارد و همان A^{-T} است.

d) در سؤال خواسته شده نشان دهیم که $\lambda_1\|w\|_2^2$ را می‌توان با اضافه کردن چند داده حذف کرد.

$$\underset{m\times 1}{Y'} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Y' و X' به شکل زیر تعریف می‌کنیم با فرض m بعدی بودن w:

$$\left.\underset{m\times m}{X'} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & \cdots & \sqrt{\lambda_1} \end{bmatrix} = \sqrt{\lambda_1}\, I \right\} \Rightarrow \|Y' - X'w\|_2^2 = \left\| \begin{bmatrix} \sqrt{\lambda_1}w_1 \\ \vdots \\ \sqrt{\lambda_1}w_m \end{bmatrix} \right\|_2^2 =$$

$$= \lambda_1\|w\|_2^2$$

بنابراین با اضافه کردن X' و Y' به X و Y داریم:

$$\underset{(m+n)\times m}{newX} = \begin{bmatrix} X' \\ X \end{bmatrix} \qquad \underset{(m+n)\times 1}{newY} = \begin{bmatrix} Y' \\ Y \end{bmatrix}$$

$$|(newX)w - newY|^2 = \left| \begin{bmatrix} X' \\ X \end{bmatrix} w - \begin{bmatrix} Y' \\ Y \end{bmatrix} \right|^2 = \|Xw - Y\|_2^2 + \|X'w - Y'\|_2^2$$

$$= \|Xw - Y\|_2^2 + \lambda_1\|w\|_2^2$$

بنابراین:

$$|Xw - Y|^2 + \lambda_1\|w\|_2^2 + \lambda_2\|w\|_1 = |newX\,w - newY|^2 + \lambda_2\|w\|_1$$

علیرضا دهقان پور فراشاه ۹۸۱۰۱۵۵۵

سؤال ۳:

$$\text{minimze} \int P(x)\log\left(\frac{P(x)}{Q(x)}\right)dx = \int P(x)\left(\log P(x) - \log Q(x)\right)dx =$$

$$KL = \int P(x)\left(\log P(x) - \log \frac{1}{(2\pi)^{n/2}}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right)dx$$

$$\frac{\partial KL}{\partial \mu} \overset{\text{لایبنیتز}}{=} \int \frac{\partial P(x)\log\left(\frac{P(x)}{Q(x)}\right)}{\partial \mu}dx = \int P(x)\left(\frac{\partial}{\partial \mu}\left(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right)dx$$

$$= \int P(x) \times -\frac{1}{2}\times(-2)\Sigma^{-1}(x-\mu)\,dx = \int P(x)\Sigma^{-1}(x-\mu)\,dx = 0$$

$$\Rightarrow \Sigma^{-1}\int xP(x)\,dx = \Sigma^{-1}\int \mu P(x)\,dx = \Sigma^{-1}\mu \underbrace{\int P(x)\,dx}_{1} = \Sigma^{-1}\mu$$

$$\Rightarrow \int xP(x)\,dx = \mu \Rightarrow \mu = E_p[X]$$

$$\frac{\partial KL}{\partial \Sigma} = \int P(x)\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}\log|\Sigma| + \frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$|\Sigma| = \sum_{j=1}^{n} \underbrace{\sigma_{ij}}_{\text{عنصر اونام}} \underbrace{|\Sigma_{ij}|}_{\text{ماتریس مجاورت}} \qquad \frac{\partial|\Sigma|}{\partial\Sigma} = \begin{bmatrix} \frac{\partial\sum\sigma_{ij}|\Sigma_{ij}|}{\partial\sigma_{11}} & \cdots & \frac{\partial\sum\sigma_{nj}|\Sigma_{nj}|}{\partial\sigma_{n1}} \\ \vdots & & \vdots \\ \frac{\partial\sum\sigma_{ij}|\Sigma_{ij}|}{\partial\sigma_{1n}} & \cdots & \frac{\partial\sum\sigma_{nj}|\Sigma_{nj}|}{\partial\sigma_{nn}} \end{bmatrix}$$

$$= \begin{bmatrix} |\Sigma_{11}| & \cdots & |\Sigma_{n1}| \\ \vdots & & \vdots \\ |\Sigma_{1n}| & & |\Sigma_{nn}| \end{bmatrix} = Adj(\Sigma) = |\Sigma|\Sigma^{-1}$$

$$\frac{\partial \log|\Sigma|}{\partial\Sigma} = \frac{1}{|\Sigma|}\begin{bmatrix} |\Sigma_{11}| & \cdots & |\Sigma_{n1}| \\ |\Sigma_{1n}| & & |\Sigma_{nn}| \end{bmatrix} = \Sigma^{-1} \Rightarrow$$

$$\frac{\partial}{\partial\Sigma}\log|\Sigma| = \Sigma^{-1}$$

درصفحه بعد به محاسبه مشتق بعدی می‌پردازیم:

$$\frac{\partial}{\partial \Sigma} (x-\mu)^T \Sigma^{-1} (x-\mu) = -\Sigma^{-1} (x-\mu)(x-\mu)^T \Sigma^{-1}$$

$$\Rightarrow \int P(x) \frac{\partial}{\partial \Sigma} \log \frac{P(x)}{Q(x)} dx = \int P(x) \left(-\frac{1}{۲} \Sigma^{-1} + \frac{1}{۲} \Sigma^{-1} (x-\mu)(x-\mu)^T \Sigma^{-1} \right) dx = 0$$

$$\Rightarrow \frac{1}{۲} \Sigma^{-1} \underbrace{\int P(x)\, dx}_{1} = \frac{1}{۲} \Sigma^{-1} \left(\int P(x) (x-\mu)(x-\mu)^T dx \right) \Sigma^{-1}$$

$$\Rightarrow \Sigma = \int (x-\mu)(x-\mu)^T P(x)\, dx = Var_p(X)$$

علیرضا دهقانپور فراشاه ۹۸۱۰۱۵۵۵

سؤال ۴:

a)

$$\frac{1}{2}\sum_{i=1}^{n}\left(w_j x_j^{(i)} - y^{(i)}\right)^2 = L$$

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} x_j^{(i)}\left(w_j x_j^{(i)} - y^{(i)}\right) = 0 \Rightarrow \sum_{i=1}^{n} w_j {x_j^{(i)}}^2 = \sum_{i=1}^{n} y^{(i)} x_j^{(i)}$$

$$\Rightarrow w_j \times \sum_{i=1}^{n} {x_j^{(i)}}^2 = \sum_{i=1}^{n} y^{(i)} x_j^{(i)} \Rightarrow w_j \times x_j^T x_j = x_j^T y$$

که $x_j = \begin{bmatrix} x_j^{(1)} \\ \vdots \\ x_j^{(n)} \end{bmatrix}$ بنابراین داریم $w_j = \dfrac{x_j^T y}{x_j^T x_j}$

---

b) طبق سؤال ۲ داریم که

$$W = (X^T X)^{-1} X y$$

$w_{m\times 1} = [w_1, \ldots, w_m]^T$

$\underset{n\times m}{X} = [x^{(0)}, \ldots, x^{(n)}]^T$

حال اگر X ماتریس orthonormal باشد داریم

$X^T X = I$ بنابراین $\Leftarrow$ $W = X y$

$$W = X y \Rightarrow \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \Rightarrow w_j = x_j^T y$$

طبق بخش قبل داشتیم که $w_j = \dfrac{x_j^T y}{x_j^T x_j}$ که چون ستون ها برهم عمود اند $x_j^T x_j = 1$ بنابراین

w_j بدست آمده در این حالت برابر حالتی است که به صورت مستقل یادگیری می‌شد.

---

c)

$w := \begin{bmatrix} w_0 \\ w_j \end{bmatrix}$ $\quad X_j^i := \begin{bmatrix} 1 \\ x_j^{(i)} \end{bmatrix}$ feature j ام از داده i ام

$\underset{2\times n}{X} := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_j^{(1)} & \cdots & \cdots & x_j^{(n)} \end{bmatrix}$

حال تابع L را به فرم ماتریسی بیان می‌کنیم:

$$L = \sum_{i=1}^{n}\left(w^T x_j^i - y^{(i)}\right)^2 = (w^T X - y^T)^T (w^T X - y^T) = (X^T w - y)(w^T X - y^T) =$$

$$= X^T w w^T X - X^T w y^T - y w^T X + y y^T$$

$$\frac{\partial L}{\partial w} = 2 w^T X X^T - y^T X^T - y^T X^T = 0 \Rightarrow w^T X X^T = y^T X^T \Rightarrow$$

$$\Rightarrow X X^T w = X y \Rightarrow w = (X X^T)^{-1} X y$$

ادامه صفحه بعد

$$XX^T = \begin{bmatrix} 1 & \cdots & 1 \\ x_j^{(1)} & \cdots & x_j^{(n)} \end{bmatrix} \begin{bmatrix} 1 & x_j^{(1)} \\ \vdots & \vdots \\ 1 & x_j^{(n)} \end{bmatrix} = \begin{bmatrix} n & \sum x_j^{(i)} \\ \sum x_j^{(i)} & \sum x_j^{2(i)} \end{bmatrix} =$$

$$= n \times \begin{bmatrix} 1 & E[x_j] \\ E[x_j] & E[x_j^2] \end{bmatrix} \Rightarrow |XX^T| = n^2 \left(E[x_j^2] - E^2[x_j]\right) = n^2 var(x_j)$$

$$\Rightarrow (XX^T)^{-1} \overset{1}{=} \frac{1}{n\, var(x_j)} \begin{bmatrix} E[x_j^2] & -E[x_j] \\ -E[x_j] & 1 \end{bmatrix}$$

$$\Rightarrow XY = \begin{bmatrix} 1 & \cdots & 1 \\ x_j^{(1)} & \cdots & x_j^{(n)} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \sum y^{(i)} \\ \sum y^{(i)} x_j^{(i)} \end{bmatrix} \overset{2}{=} n \times \begin{bmatrix} E[y] \\ E[y x_j] \end{bmatrix}$$

1×2

$$\Rightarrow \begin{bmatrix} w_0 \\ w_j \end{bmatrix} = (XX^T)^{-1} \times Y = \frac{1}{Var(x_j)} \begin{bmatrix} E[x_j^2]E[y] - E[x_j]E[x_j y] \\ -E[x_j]E[y] + E[x_j y] \end{bmatrix} =$$

$$\Rightarrow w_j = \frac{E[x_j y] - E[x_j]E[y]}{Var(x_j)} = \frac{Cov(x_j, y)}{Var(x_j)}$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{n} [1 \;\; x_j^{(i)}] \left(w_j x_j^{(i)} + w_0 - y^{(i)} \right) = 0 \Rightarrow$$

$$\sum_{i=1}^{n} -w_j x_j^{(i)} + y^{(i)} = \sum_{i=1}^{n} w_0 = n w_0 \Rightarrow$$

$$\Rightarrow w_0 = \frac{1}{n} \sum y^{(i)} - \frac{1}{n} \sum x_j^{(i)} w_j = E[y] - w_j E[x_j]$$

$$\Rightarrow w_0 = E[y] - w_j E[x_j]$$