تمرین ML

نوری
امیرحسین باقری 98105621

سوال اول

بارت اول: گزینه c صحیح است.

دقت کنید که زیاد بودن ضریب یکی از feature ها به معنای مهم بودن آن است. (رد گزینه b) امّا با توجه به اینکه درباره توزیع داده‌ها correlation آنها اطلاعی نمی‌توان درباره حذف نظر داد ← ازاین‌رو است که feature ها با correlation زیادی دارند. آن‌گاه فیچی دارای ضریب بزرگ با علامت مخالف باشد. و این دو همدیگر را خنثی کنند.

با پرانتز درباره retained کردن آن نمی‌توان نظر داد ← (a) ⇐ بنابراین c صحیح است.

بارت دوم:

a) درست است.

زمانی که Bias زیاد است. به معنای آن است که مدل underfitting دارد و نمی‌تواند توزیع را به خوبی مدل کند و افزایش تعداد training set ها تأثیری در این مورد حل نمی‌گذارد. مثلا ← $f(x)=x^2+x^3$ را با خطی مدل کنیم. Poly از درجه ۲ حل کنیم.

$\hat{f}(x)=ax+b$ ⇐ هرچه تعداد data زیاد شود باز هم $\hat{f}(x)=ax+b$ نمی‌تواند $f(x)=x^2+x^3$ را مدل کند.

b) غلط است.

زیرا اگر مدل overfit شود، آنگاه می‌توان ادر داده‌های train را کم کرد. امّا می‌توان test error را کاهش داد. ←

overfitted model → increase complexity

still overfitted with lower training error but higher test error.

c) ← غلط است.

اگر مدل underfit باشد نیاز داریم تا complexity مدل را افزایش دهیم. و این افزایش به کاهش test error و training error منجر می‌شود.

underfitted model → increase complexity → decrease test error, training error

d) ← غلط است ← زمانی که training error زیاد است. بدان معناست که باید complexity مدل افزایش یابد، تا Bias کاهش یابد ← باید از یک Polynomial regression پیچیده‌تر (با درجه بیشتر) استفاده کرد. نه یک Linear regression. (دقت کنید که $\phi(x)$ را یکی در نظر گرفتیم به Poly با درجه بیشتر برسیم امّا ثابت فرض کرده‌ایم.)

سوال ۲

a) ثابت کنید که β همان رابطه است.

Residuals are $\varepsilon = Y - X\beta$

$$\text{error} = \sum_N \varepsilon_i^2 = [\varepsilon_1 \cdots \varepsilon_n] \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \varepsilon^T \varepsilon$$

we want to minimize $\varepsilon^T\varepsilon = (Y - X\beta)^T (Y - X\beta)$

$$\frac{\partial (Y - X\beta)^T (Y - X\beta)}{\partial \beta} = -2X^T(Y - X\beta) \Rightarrow X^T(Y - X\beta) = 0 \Rightarrow X^T X\beta = X^T Y$$

$$\Rightarrow \beta = (X^T X)^{-1} X^T Y$$

b) with L_2 regularization $\Rightarrow$ $\text{error} = \sum_N \varepsilon_i^2 + \lambda \sum \beta_i^2$

$$\text{error} = \varepsilon^T\varepsilon + \lambda \beta^T\beta \Rightarrow (Y - X\beta)^T (Y - X\beta) + \frac{\lambda}{2}\beta^T\beta$$

$$\frac{\partial \text{error}}{\partial \beta} = -2X^T(Y - X\beta) + 2\beta\frac{\lambda}{2} = 0 \Rightarrow X^T X\beta + \lambda\beta = X^T Y$$

$$(X^T X + \lambda I)\beta = X^T Y = \boxed{\beta = (X^T X + \lambda I)^{-1} X^T Y}$$

$$w^* = (X^T X + \lambda I)^{-1} X^T Y$$

if $\lambda = 0 \longrightarrow w^* = (X^T X)^{-1} X^T Y$

c) ثابت کنید که $\Sigma X = XF$ است حال داریم $\leftarrow$

$$\Sigma X = XF \Rightarrow X^T\Sigma^T = F^T X^T \Rightarrow X^T\Sigma = F^T X^T$$

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

حال در w^*_{new} عبارت بالا را $X^T\Sigma^{-1}$ جایگزین می‌کنیم.

$$w^*_{new} = (X^T\Sigma^{-1}X)^{-1} X^T\Sigma^{-1}Y = \underbrace{(F^{-T}X^T X)^{-1} F^{-T}}_{(X^T X)^{-1} F^T} X^T Y = (X^T X)^{-1} X^T Y$$

$\Rightarrow$ if $\Sigma X = XF$ and F Not singular $\longrightarrow$ $w^*_{new} = w^*_{old} \longrightarrow$

if $\Sigma X = XF \Rightarrow w^*_{new} = w^*_{old}$ ✓

ادامه سوال ۲ C

$w^*_{new} = w^*_{old}$

فرض کنید ←

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

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

$$\Rightarrow X^T = \left[(X^T\Sigma^{-1}X)(X^TX)^{-1}\right]^{-1}X^T\Sigma^{-1} \xRightarrow{\times\Sigma}$$

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

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

Σ متقارن

ماتریس $(X^T\Sigma^{-1}X)$ حکوس پذیر است.

ماتریس $(X^TX)^{-1}$ نیز حکوس پذیر است.

$\Leftarrow$ $\left[(X^T\Sigma^{-1}X)(X^TX)^{-1}\right]^{-1}$ حکوس پذیر است.

$$F^T = \left[(X^T\Sigma^{-1}X)(X^TX)^{-1}\right]^{-1} \Leftarrow$$

$$F = \left(\left[(X^T\Sigma^{-1}X)(X^TX)^{-1}\right]^{-1}\right)^T$$

F یک ماتریس non singular است.

بنابراین

if $w^*_{new} = w^*_{old} \Rightarrow$ we found F which is non singular matrix and

$$\Sigma X = XF \checkmark$$

if $\Sigma X = XF$ (F non singular) $\Rightarrow$ $w^*_{new} = w^*_{old} \Rightarrow$

$w^*_{old} = w^*_{new} \iff \Sigma X = XF$ (F is non singular).

Proof Done ✓

فرضیه بعدی یک توضیح اضافه می‌دهیم.

$$(X^TX)^{-1}X^TY = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$ دقت کنید که گفته شده:

$$\underbrace{\left[(X^TX)^{-1}X^T - (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right]}_{C}Y = 0 \Rightarrow CY = 0 \quad \Leftarrow$$

دقت کنید که به ازای هر Y_0 رابطه بالا برقرار است زیرا Y یک توزیع آماری دارد

$$\forall Y \; CY = 0 \Longrightarrow \boxed{C = 0} \quad \Leftarrow$$

قسمت d سوال ۲

اگر w دارای ابعاد D باشد ←

باید به loss term یک عبارت $\lambda_1 w_i^2$ اضافه گردد.

⟸ برای اینکار ← به Y سطرهای D تا 0 را اضافه می‌کنیم و آنرا صفر می‌گذاریم.

$$Y' = \begin{bmatrix} \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} N \\ D \end{matrix}$$

برای ستون X هارا به:

$$X_w = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \sqrt{\lambda_1} & \\ & & \ddots \\ & & & \sqrt{\lambda_1} \end{bmatrix}_{D \times D}$$

$$\Longrightarrow X' = \begin{bmatrix} X_w \\ X \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_1} \\ & X & \end{bmatrix} \begin{matrix} \\ \\ \\ \rightarrow N+D \end{matrix}$$

حال داریم:

$$error = \sum |Y_i' - X_i' w|^2 + \lambda_2 \sum |w_i|$$

$$= error = \sum_{i}^{D} |Y_i' - X_i' w|^2 + \sum_{D+1}^{N+D} |Y_i' - X_i' w|^2 + \lambda_2 \sum |w_i|$$

$$error = \sum |X_i' w|^2 + \sum |Y_i - X_i w|^2 + \lambda_2 \sum |w_i| \Rightarrow$$

$$\sum (\sqrt{\lambda_1} w_i)^2 + \sum |Y_i - X_i w|^2 + \lambda_2 \sum |w_i| = \sum \lambda_1 |w_i|^2 + \sum |Y_i - X_i w|^2 + \lambda_2 \sum |w_i|$$

$$error = (Y - Xw)^T (Y - Xw) + \lambda_1 w^T w + \lambda_2 \sum |w_i|$$

$$error = (Y' - X'w)^T (Y' - X'w) + \lambda_2 \sum |w_i| \quad \checkmark$$

سوال ۳

روش اول

قسمت اول

$$f(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\text{minimize } KL = \text{minimize} \underbrace{\int p(x)\lg p(x)\,dx}_{\text{independent of } \mu \text{ and } \Sigma} - \int p(x)\lg q(x)\,dx$$

$$\text{Minimization of } KL = \text{maximization} \int p(x)\lg q(x)\,dx = \text{minimize}$$

$$Y = \int p(x)\left[\lg\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right) - \frac{1}{2}\exp\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right]dx = Y$$

$$\text{maximize for } \mu \Rightarrow \frac{\partial Y}{\partial \mu} = -\frac{1}{2}\int 2\Sigma^{-1}(x-\mu)\,p(x) = 0$$

$$\Rightarrow \int (x-\mu)\,p(x) = 0 \Rightarrow \int x\,p(x)\,dx = \mu\underbrace{\int p(x)\,dx}_{1} \Rightarrow \boxed{\mu = E_p(x)}$$

قسمت اول

قسمت دوم

ابتدا توجه کنید از مرتبه اول می دانیم $\Sigma^T = \Sigma$ ، $\frac{\partial}{\partial A}Y^TAY = YY^T$ ، $\frac{1}{|A|} = |A^{-1}|$

$\frac{\partial}{\partial A}\lg|A| = (A^{-1})^T$ ← اثبات آخر سوال ۳

$$KL \text{ maximization} = \text{maximize} \int p(x)\lg\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right) - \frac{1}{2}\int \exp\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx$$

$$\Rightarrow \frac{\partial Y}{\partial \Sigma} = \int p(x)(\lg 2\pi)^{p/2}\,dx + \frac{1}{2}\int p(x)\lg|\Sigma^{-1}| - \frac{1}{2}\int p(x)(x-\mu)^T\Sigma^{-1}(x-\mu)\,dx$$

$$\frac{\partial Y}{\partial \Sigma^{-1}} = \int \frac{\partial}{\partial \Sigma^{-1}} p(x)\lg(2\pi)^{p/2} + \frac{1}{2}\int p(x)(\Sigma^{-1})^{-1}\,dx - \frac{1}{2}\int p(x)(x-\mu)(x-\mu)^T\,dx$$

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

$$\Rightarrow \Sigma = E\{(x-\mu)(x-\mu)^T\} = Var_{p(x)}$$

ادامه دوم

به صورت کلی

اثبات می‌کنیم که minimize کردن KL برابر maximum likelihood است.

ابتدا برای یک تابع convex حالت $\ln$ داریم $\Rightarrow f(\int g(x)) \geq \ln(\int f(g(x)))$

می‌دانیم $\ln$ کانوکس است بنابراین

$$KL(P\|q) = -\int P(x) \ln \frac{q(x)}{P(x)} dx \geq -\ln \int \underbrace{P(x) \frac{q(x)}{P(x)} dx}_{1} \geq -\ln 1$$

$$KL(P\|q) = -\int P(x) \ln \frac{q(x)}{P(x)} \geq 0$$

تابع $KL(P\|q) \geq 0$ است $\Leftarrow$ تابع $-\ln(x)$ تنها در اصل است.

بنابراین تنها زمانی KL صفر است که $q(x)=P(x)$ باشد.

حال چون توزیع $P(x)$ را می‌دانیم داریم $P(x) \approx q(x|\theta)$

θ را می‌دانیم پس ثابت است.

$$KL(P\|q) = \sum_N -\ln\left(\frac{Q(x|\theta)}{P(x_i)}\right) P(x_i) \Rightarrow \frac{\partial KL}{\partial(\hat{P}_i)} = \sum_N \frac{\partial}{\partial P(x)} \left(-\ln q(x|\theta) + \ln P(x_i)\right)$$

$$\text{minimize} \sum_N -\ln q(x|\theta) = \text{maximize} \sum \ln(Q|\theta) \Rightarrow \text{maximize} \prod_N Q(x|\theta)$$

که معادل $\Leftarrow$ maximum likelihood است $\Leftarrow$ از maximum likelihood می‌دانیم.

* از تخمین ML برآورد می‌کنیم.

$$\mu = E_p(x)$$

$$\Sigma = \text{Var}(x)$$

$$\frac{\partial \ln|A|}{\partial A} = (A^{-1})^T$$ — (ب)

$$\det(A) = \sum_j (-1)^{i+j} A_{ij} M_{ij}$$

$$\frac{\partial}{\partial A_{ij}} \ln|A| = \frac{C_{ij}}{|A|} \rightsquigarrow \frac{\partial \ln|A|}{\partial A_{ij}} = \frac{C}{|A|}$$

$$A^{-1} = \frac{C^T}{|A|} \Rightarrow (A^{-1})^T = \frac{C}{|A|} = (A^{-1})^T$$ ✓

سوال 4 ام

a)

we apply only on col jth

$$y_i = w_i x_{ij} + \varepsilon_i \sim N(0, \Sigma) \qquad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\rightarrow$ shows the col jth *

$$Y = w_i X_j + \varepsilon \Rightarrow \text{error} = \sum_N \varepsilon_i^2 = \varepsilon^T \varepsilon$$

minimiz $(Y - w_j X_j)^T (Y - w_j X_j) \rightsquigarrow \frac{\partial e}{\partial w_j} = -2X^T(Y - w_j X_j) = 0$

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

که X_j نشان دهنده ستون j ام X است.

b)

$$w^* = (X^T X)^{-1} X^T Y$$

$\boxed{X^T X = I}$ ستند orthogonal از آنجا که در صورت سوال گفته شده که ستون های X

$$w^* = X^T Y \Rightarrow w_j^* = (X^T Y)_j \Rightarrow w_j^* = X_j^T Y \quad (I)$$

ستون j ام

حال از بخش الف استفاده می‌کنیم $\leftarrow$ داریم feature هر روی train

$$w_j^* = \frac{X_j^T y}{X_j^T X_j} \qquad X_j^T X_j = 1 \quad (II)$$

$\hookrightarrow$ orthogonality.

$(I), (II) \Rightarrow w_j^* = X_j^T y$

اگر X ستون های orthogonal است با توجه به نتیجه حاصل از train روی هر feature

با حلت train روی کل feature ها یکی است.

در واقع چون X orthogonal است $\Leftarrow$ Projection هر ستون روی ستون دیگر صفر است.

$\Leftarrow$ [redacted]

ادامه 4

ج

$$error = \sum_{i=1}^{N} (y_i - w_j x_{ij} - w_0)^2 \Rightarrow \frac{\partial error}{\partial w_0} = 0$$

$$\frac{\partial error}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{N} (y_i - w_j x_{ij} - w_0) = 0 \Rightarrow \sum_{i=1}^{N} w_0 = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} w_j x_{ij}$$

$$\Rightarrow N w_0 = \sum_{i=1}^{N} y_i - w_j \sum_{i=1}^{N} x_{ij} \Rightarrow w_0 = \frac{\sum_{i=1}^{N} y_i}{N} - w_j \frac{\sum_{i=1}^{N} x_{ij}}{N} \Rightarrow$$

$$w_0 = E(y) - w_j E[x_j] = E(y) - w_j E(x_j) \checkmark \rightarrow (I)$$

$x_j \leftarrow$ ستون ام است.

میانگین ستون ام

$$\frac{\partial error}{\partial w_j} = \sum_{i=1}^{N} 2 x_{ij} (y_i - w_j x_{ij} - w_0) = 0 \Rightarrow \sum_{i=1}^{N} w_j x_{ij}^2 = \sum_{i=1}^{N} x_{ij} y_i - \sum_{i=1}^{N} w_0 x_{ij}$$

$$\Rightarrow w_j \sum_{i=1}^{N} x_{ij}^2 = \sum_{i=1}^{N} x_{ij} y_i - w_0 \times N \times E(x_j) \Rightarrow$$ (I) subsitud the w_0 in this

$$\Rightarrow w_j \sum_{i=1}^{N} x_{ij}^2 = \sum_{i=1}^{N} x_{ij} y_i - [E(y) - w_j E(x_j)] N E(x_j) \Rightarrow$$

$$w_j \left(\sum_{i=1}^{N} x_{ij}^2 - N E(x_j)^2 \right) = \sum_{i=1}^{N} x_{ij} * y_i - N E(y) E(x_j) \Rightarrow$$ تقسیم بر N

$$w_j \left(\frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 - E(x_j)^2 \right) = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} y_i) - E(y) E(x_j) \Rightarrow$$

$\left(\frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 - E(x_j)^2 \right) = var(x_j)$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 = E(x_j^2) \Rightarrow E(x_j)^2 - E(x_j)^2 = var(x_j)$$

$$\frac{1}{N} \sum_{i=1}^{N} (x_{ij} y_i) = E(x_j y) \Rightarrow E(x_j y) - E(y) E(x_j) = cov(x_j, y)$$

$$\Rightarrow \boxed{w_j = \frac{cov(x_j, y)}{Var(x_j)}} \checkmark$$

Done.