HW1 writing 王若飞 2021533135.

No.

Date. / /

1 1) prove X_i's covariance is a semi positive definite matrix

covariance $\Sigma = E[(X-\bar{X})(X-\bar{X})^T]$ ← ($\bar{X}=M_i$ is the average of X_i)

then for $\forall$ real vector u. we can have

$$u^T\Sigma u = u^T E[(X-\bar{X})(X-\bar{X})^T]u$$
$$= E[u^T(X-\bar{X})(X-\bar{X})^T u]$$

we can see $u^T(X-\bar{X}) = G_i = ((X-\bar{X})^T u)^T$

$\therefore$ we can have

$$u^T\Sigma u = E[G_i G_i^T] = E[G_i^2] = E[\|u^T(X-\bar{X})\|_2^2]$$

≥ 0

so finally we get it.

2) Because $X_i \sim \mathcal{N}(M, \Sigma)$ which is a mul normal distribution

$\therefore$ we can directly get it PDF

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

then we use log-likelihood to get the $\ell(M,\Sigma)$

$$\Rightarrow \ell(u,\Sigma) = \sum_{i=1}^{N} \log f(x_i) = \sum_{i=1}^{N} \log\left[\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_i-M)^T\Sigma^{-1}(x_i-M)\right)\right]$$
$$= \sum_{i=1}^{N} \log \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} - \frac{1}{2}\sum_{i=1}^{N}(x_i-M)^T\Sigma^{-1}(x_i-M)$$

then we can use it to get the MLE of u.

$$\Rightarrow \frac{\partial \ell(u,\Sigma)}{\partial u} = \frac{\partial\left[\sum_{i=1}^{N}\log\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} - \frac{1}{2}\sum_{i=1}^{N}(x_i-M)^T\Sigma^{-1}(x_i-M)\right]}{\partial u} = -\frac{1}{2}\frac{\partial \sum_{i=1}^{N}(x_i-M)^T\Sigma^{-1}(x_i-M)}{\partial u}$$

not corresponding

$$= -\frac{1}{2}\cdot(-2)\cdot\sum_{i=1}^{N}\Sigma^{-1}(x_i-M)$$
$$= \sum_{i=1}^{N}\Sigma^{-1}(x_i-M)$$

Let it equal to 0

$$\therefore \sum_{i=1}^{N}\Sigma^{-1}(x_i-M) = 0 \qquad \therefore \hat{u} = \frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\therefore \sum_{i=1}^{N}(x_i-N) = 0$$

No.

Date. / /

3) ∵ $\hat{\theta}$ is an unbiased estimator of θ

∴ $E[\hat{\theta}] = \theta$

∵ $Var(X) = E[X^2] - E[X]^2$

∴ $Var(\hat{\theta}) = E[\hat{\theta}^2] - E[\hat{\theta}]^2$

$= E[\hat{\theta}^2] - \theta^2 > 0$

∴ $E[\hat{\theta}^2] > \theta^2$

∴ $\hat{\theta}^2$ is not an unbiased estimator of θ^2

王芸涵　2021533135

∵ multiply　No.　calculate $\frac{\partial L(a\cdot b)}{\partial a}$ and $\frac{\partial L(a\cdot b)}{\partial b}$ are too complicated to

MLE problem

2 1) The likelihood function is $L(a\cdot b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y-aX-b)^2\right)$ ✓

then we can use log-likelihood to get the simplified expression so that we can ~~get~~ estimate a, b more easily ⇒ ∵ add

$$\Rightarrow \underset{a\cdot b}{\operatorname{argmax}}\, \ell(a\cdot b) \triangleq \underset{a\cdot b}{\operatorname{argmax}}\, L(a\cdot b)$$

then
$$\ell(a\cdot b) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i-ax_i-b)^2\right)$$
$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i-ax_i-b)^2\right)$$
$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b)^2$$

Thus, we can get the correct a, b to maximize $\ell(a\cdot b)$ through $\frac{\partial \ell(a\cdot b)}{\partial a}$ or $\frac{\partial \ell(a\cdot b)}{\partial b}$, and this way's effect is equal to maximize $L(a\cdot b)$ ✓

2) ① $$\frac{\partial \ell(a\cdot b)}{\partial a} = -\frac{1}{2\sigma^2}\frac{\partial \sum_{i=1}^{n}(y_i-ax_i-b)^2}{\partial a} = -\frac{1}{2\sigma^2}(-2)\sum_{i=1}^{n}x_i(y_i-ax_i-b)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i-ax_i-b)$$

Let $\frac{\partial \ell(a\cdot b)}{\partial a}$ be equal to 0

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i-ax_i-b) = 0$$
$$\sum_{i=1}^{n}x_i(y_i-b) = \sum_{i=1}^{n}ax_i^2$$
$$\frac{\sum_{i=1}^{n}x_i(y_i-b)}{\sum_{i=1}^{n}x_i^2} = \hat{a}$$

② $$\frac{\partial \ell(a\cdot b)}{\partial b} = -\frac{1}{2\sigma^2}\frac{\partial \sum_{i=1}^{n}(y_i-ax_i-b)^2}{\partial b} = -\frac{1}{2\sigma^2}\cdot(-2)\sum_{i=1}^{n}(y_i-ax_i-b)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b)$$

Let $\frac{\partial \ell(a\cdot b)}{\partial b}$ be equal to 0

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b) = 0 \qquad \frac{\sum_{i=1}^{n}(y_i-ax_i)}{n} = \hat{b}$$

$$\Rightarrow \hat{a} = \frac{\sum_{i=1}^{n}x_i(y_i-b)}{\sum_{i=1}^{n}x_i^2}, \quad \hat{b} = \frac{\sum_{i=1}^{n}(y_i-ax_i)}{n}$$

Not simplify

$$\Rightarrow ① \hat{b} = \frac{1}{n}\sum_{i=1}^{n}y_i - \frac{\hat{a}}{n}\sum_{i=1}^{n}x_i$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x} = \bar{y} - \bar{x}\cdot\frac{\sum_{i=1}^{n}(y_i-\bar{y})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}$$
$$② \hat{a} = \frac{\sum_{i=1}^{n}x_i(y_i-\bar{y}+\hat{a}\bar{x})}{\sum_{i=1}^{n}x_i^2}$$
$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}$$

simplify

No.

Date. / /

3) $f(x) = ax + b$

let we substitute $\hat{a}$. $\hat{b}$ in this expression

$f(x) = \hat{a}x + \hat{b}$

and then substitute $\bar{x}$ in it.

$f(\bar{x}) = \hat{a}\bar{x} + \hat{b}$

$= \frac{\hat{a}}{n}\sum_{i=1}^{n} x_i + \hat{b}$

$= \frac{\hat{a}}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i)$

$= \frac{\hat{a}}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\hat{a}}{n}\sum_{i=1}^{n} x_i$

$= \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ $\quad \therefore \boxed{f(\bar{x}) = \bar{y}}$

Therefore, based on the result in (b): $\hat{a}$, $\hat{b}$. the learned linear model $f(x) = ax + b$, always passes through the point $(\bar{x}, \bar{y})$.

REMEMBER · MEMORY

고분자 2021-33135

No.
Date. / /

3. D "Linear" regards that the input features and the outputs have a linear relation.

2) From what we have learned in class. we define

$$\hat{\beta}^{ridge} = \arg\min_{\beta} PRSS(\lambda,\beta) = \arg\min_{\beta} \|y-X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\therefore PRSS(\lambda,\beta) = \|y-X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$= (y-X\beta)^T(y-X\beta) + \lambda\beta^T\beta$$

$$= y^Ty - \beta^TX^Ty - y^TX\beta + \beta^TX^TX\beta + \lambda\beta^T\beta$$

to minimise $PRSS(\lambda,\beta)$ we take the derivative of it

$$\frac{\partial PRSS(\lambda,\beta)}{\partial\beta} = -X^Ty - X^Ty + 2X^TX\beta + 2\lambda I_p\beta$$

$$= -2X^Ty + 2(X^TX+\lambda I_p)\beta$$

In this expression, λ is the parameter which we have determine before this step

Let it be equal to $\frac{\partial PRSS(\lambda,\beta)}{\partial\beta} = 0$

$$\Rightarrow \hat{\beta}^{ridge} = (X^TX+\lambda I_p)^{-1}X^Ty = \beta_{ridge} \quad \checkmark$$

Thus, we simply get this expression. but there is also a problem we need to prove. that is $(X^TX+\lambda I_p)^{-1}$ should be invertible, given $\lambda > 0$

$$\Rightarrow X^TX+\lambda I_p = ((UDV^T)^T(UDV^T)) + \lambda I_p$$

SVD decomposition. $X = UDV^T$, $V^TV = U^TU = I_p$

$$= VDU^TUDV^T + \lambda I_p$$

$$= VD^2V^T + \lambda VI_pV^T$$

$$= V(D^2+\lambda I_p)V^T$$

D^2: this is semi positive definite

λI_p: $\lambda > 0$ and I_p is a Identity matrix

$\therefore D^2+\lambda I_p$ is positive definite

$\therefore (X^TX+\lambda I_p)$ is invertible

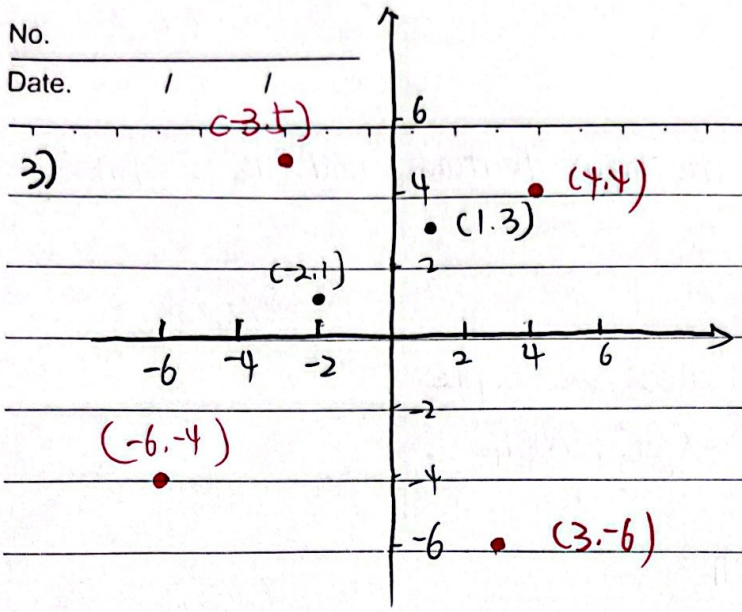
$\therefore$ Finally, we get it $\checkmark$

Wengu文谷

No.

Date. / /

3)



Therefore, from the graph above, we can know the given data set is not linear separable. We can't use one straight line to separate these two labels from each other.

Instead, we can use a circle to sperate them. which is Multivariate nonlinear ✓