王芸（?） 2021533135

HW2

No.

Date. / /

1. a) proof any norm $f: R^n \to R$ is convex

$\Rightarrow$ P-norm : $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$, $1 \le p$

$\Rightarrow$ convex : $\forall v, w \in dom\, f$, $0 \le \theta \le 1$: $f(\theta v + (1-\theta)w) \le \theta f(v) + (1-\theta) f(w)$

$\Rightarrow$ property of P-norm : ① $\forall v \in f$: $\|v\|_p \ge 0$ and $\|v\|_p = 0 \Leftrightarrow v = 0$ (positive)

$p \ge 1$ ② $\forall v \in f; \lambda \in R$: $|\lambda| \|v\|_p = \|\lambda v\|_p$ (absolutely scaleable)

③ $\forall w, v \in f$: $\|v+w\|_p \le \|v\|_p + \|w\|_p$ (triangle inequality)

Therefore from property ②, ③ and the definition of convex, then we can have

$$\|\lambda v + (1-\lambda) w\|_p \overset{③}{\le} \|\lambda v\|_p + \|(1-\lambda) w\|_p \overset{②}{\le} |\lambda| \|v\|_p + |1-\lambda| \|w\|_p$$

$\Rightarrow$ when $\lambda \in [0, 1]$, we can have

$$\|\lambda v + (1-\lambda) w\|_p \underset{\text{definition of convex}}{\le} \lambda \|v\|_p + (1-\lambda) \|w\|_p$$

$\therefore$ we get it, that is any norm $f: R^n \to R$ is convex.

b) $f(x_1, x_2) = \frac{x_1^2}{x_2}$ on $R \times R > 0$

Hessian $\nabla^2 f(x_1, x_2) = \begin{bmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial^2 x_1} , & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} , & \frac{\partial^2 f(x_1,x_2)}{\partial^2 x_2} \end{bmatrix} = \begin{bmatrix} \frac{2}{x_2} , & -\frac{2x_1}{x_2^2} \\ \frac{-2x_1}{x_2^2} , & \frac{2x_1^2}{x_2^3} \end{bmatrix}$

$$= \frac{2}{x_2} \begin{bmatrix} 1 , & -\frac{x_1}{x_2} \\ -\frac{x_1}{x_2} , & \frac{x_1^2}{x_2^2} \end{bmatrix} = \frac{2}{x_2} \begin{bmatrix} 1 \\ -\frac{x_1}{x_2} \end{bmatrix} \begin{bmatrix} 1 , & -\frac{x_1}{x_2} \end{bmatrix} \ge 0$$

$\therefore f(x_1, x_2) = \frac{x_1^2}{x_2}$ on $R \times R > 0$ is convex.

c) $f(x_1, x_2) = \frac{x_1}{x_2}$ on $R^2 > 0$

Hessian $\nabla^2 f(x_1, x_2) = \begin{bmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial^2 x_1} , & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} , & \frac{\partial^2 f(x_1,x_2)}{\partial^2 x_2} \end{bmatrix} = \begin{bmatrix} 0 , & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} , & \frac{2x_1}{x_2^3} \end{bmatrix} = H$

$\therefore f(x_1, x_2) = \frac{x_1}{x_2}$ on $R \times R > 0$ is not convex or concave, the reason is as follows.

$$\lambda I - H = \begin{bmatrix} \lambda , & +\frac{1}{x_2^2} \\ +\frac{1}{x_2^2} , & \lambda - \frac{2x_1}{x_2^3} \end{bmatrix} \qquad \therefore \lambda\left(\lambda - \frac{2x_1}{x_2^3}\right) - \frac{1}{x_2^4} = 0$$

No.

Date. / /

$\lambda^2 - \frac{2x_1}{x_2^3}\lambda - \frac{1}{x_2^4} = 0 \qquad \Delta = b^2 - 4ac = \sqrt{\frac{4x_1^2}{x_2^6} + \frac{4}{x_2^4}} \geq 0$

$\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a} \qquad \therefore \lambda = \frac{\frac{2x_1}{x_2^3} \pm \sqrt{\frac{4x_1^2}{x_2^6} + \frac{4}{x_2^4}}}{2}$

$= \frac{x_1}{x_2^3} \pm \sqrt{\frac{x_1^2 + x_2^2}{x_2^6}} = \frac{1}{x_2^3}\left(x_1 \pm \sqrt{x_1^2 + x_2^2}\right)$

$\therefore \exists\ \lambda_1 < 0.\ \lambda_2 > 0$ condition $\therefore$ H is neither positive semi-definite nor negative semi-definite

$\therefore$ We get it.

No.

Date. /

$x > 0$

(d) $f(x) = x\log x$ is strictly convex

$$f'(x) = x\cdot\frac{1}{x} + \log x = 1+\log x \quad \uparrow$$

$$f''(x) = \frac{1}{x},\ x>0 \quad \therefore f''(x) > 0$$

$\Rightarrow$ according to Jensen's inequality, we have

$$\sum_i \alpha_i f(x_i) \geq f\left(\sum_i \alpha_i x_i\right)$$

for $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$. Set $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ and $x_i = \frac{a_i}{b_i}$, substitute, we have

$$\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\log\frac{a_i}{b_i} = \sum_i \frac{a_i}{\sum_j b_j}\log\frac{a_i}{b_i}$$

$$\geq$$

$$\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\log\sum_i\frac{b_i}{\sum_j b_j}\frac{a_i}{b_i} = \sum_i\frac{a_i}{\sum_j b_j}\log\sum_i\frac{a_i}{\sum_j b_j}$$

$$\therefore \sum_i a_i\log\frac{a_i}{b_i} \geq \left(\sum_i a_i\right)\log\frac{\sum_i a_i}{\sum_j b_j}$$

$\therefore$ we get it.

REMEMBER · MEMORY

王芸飞　2021533135

No.

Date.　/　/

2 a) from the def of $p(y_c^i=1|x^i;\beta)=\frac{e^{\beta_c^T x_i}}{\sum_{c'} e^{\beta_{c'}^T x_i}}$

$\therefore \sum_{c=1}^{K} p(y_c^i=1|x_i;\beta)=1$ ✓

$\therefore \Rightarrow \ell(\beta)=\ln\prod_{i=1}^{n} p(y_t|x^i;\beta)$　　t means the true class for x^i

$\Downarrow$

$y_t^i=1$, and other $c\neq t$, $y_c^i=0$ ✓

$$=\sum_{i=1}^{n}\sum_{c=1}^{k} y_c^i \ln p(y_c^i|x^i;\beta)$$

$$=\sum_{i=1}^{n}\sum_{c=1}^{k} y_c^i \ln\frac{e^{\beta_c^T x_i}}{\sum_{c'} e^{\beta_{c'}^T x_i}}$$

$$=\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i\left(\beta_c^T x_i-\ln\sum_{c'} e^{\beta_{c'}^T x^i}\right)\right]$$

$$=\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^T x_i)-y_c^i\ln\left(\sum_{c'} e^{\beta_{c'}^T x^i}\right)\right]$$

$\therefore$ we get it

b) $\nabla_{\beta_1}\ell(\beta)=\nabla_{\beta_1}\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^T x^i)-y_c^i\ln\left(\sum_{c'} e^{\beta_{c'}^T x^i}\right)\right]$

$\therefore$ take the derivative of β_1

(1) $\nabla_{\beta_1}\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i(\beta_c^T x^i)\right]$　　$\therefore$ when $c\neq 1$, other terms can be seen as constant

$$=\sum_{i=1}^{n} y_1^i x^i$$

(2) $\nabla_{\beta_1}\sum_{i=1}^{n}\sum_{c=1}^{k}\left[y_c^i\ln\left(\sum_{c'} e^{\beta_{c'}^T x^i}\right)\right]$

$=\nabla_{\beta_1}\sum_{i=1}^{n}\left(\sum_{c=1}^{k} y_c^i\right)\ln\left(\sum_{c'} e^{\beta_{c'}^T x_i}\right)$　　means $y_i^i=1$, others equal to 0

$$=\nabla_{\beta_1}\sum_{i=1}^{n}\ln\left(\sum_{c'} e^{\beta_{c'}^T x_i}\right)$$

$$=\sum_{i=1}^{n}\frac{e^{\beta_1^T x^i}}{\sum_{c'} e^{\beta_{c'}^T x^i}}\cdot x^i=\sum_{i=1}^{n} p(y_1^i=1|x^i;\beta)\cdot x^i$$

$\therefore \nabla_{\beta_1}\ell(\beta)=(1)-(2)$

$$=\sum_{i=1}^{n} y_1^i x^i-\sum_{i=1}^{n} p(y_1^i=1|x^i;\beta)\cdot x^i=\sum_{i=1}^{n} x^i\left(y_1^i-p(y_1^i=1|x^i;\beta)\right)$$

王冀?? 2025-33135.

No.
Date. / /

3 a) from the question. we already have $X \sim Expo(\lambda)$. $\lambda \sim Gamma(\alpha, \beta)$
and $D = \{X_1, X_2 \cdots X_n\}$, in which they are i.i.di variables

$\therefore$ from Bayes rule, we can have

$$P(\lambda|D) = \frac{P(D|\lambda)P(\lambda)}{P(D)} \overset{i.i.di}{=\!=} \frac{\prod_i^n P(X_i|\lambda)P(\lambda)}{P(D)}$$

and because $P(D)$ can be seen as constant which has no λ parameter

$$\therefore P(\lambda|D) \propto \prod_i^n P(X_i|\lambda)P(\lambda)$$

$$\propto \lambda^n \cdot e^{-\lambda(X_1+X_2+\dots+X_n)} \cdot \lambda^{4-1} e^{-\lambda\beta}$$ in this step, we drop $\frac{\beta^4}{\Gamma(4)}$. because it is constant

$$\propto \underline{\lambda^{n+4-1} \cdot e^{-\lambda(\sum_{i=1}^n x_i+\beta)}}$$

this form is similar with gamma distribution without multiplier

$$\therefore \quad P(\lambda|D) \propto Gamma(n+4, \sum_{i=1}^n x_i+\beta)$$

$\therefore$ we get it.

b) MAP method

$$\Rightarrow \log P(\lambda|D) \propto \underline{-\lambda(\sum_{i=1}^n x_i+\beta) + (n+4-1)\log\lambda}$$

$$\nabla_\lambda \log P(\lambda|D) \underset{\text{approximately}}{\propto} \nabla_\lambda\left[(-\lambda(\sum_{i=1}^n x_i+\beta)) + (n+4-1)\log\lambda\right]$$

$$= -(\sum_{i=1}^n x_i+\beta) + \frac{n+4-1}{\lambda}$$

Let it be equal to 0

$$\therefore \quad \sum_{i=1}^n x_i+\beta = \frac{n+4-1}{\lambda}$$

$$\hat{\lambda}^{MAP} = \frac{n+4-1}{\sum_{i=1}^n x_i+\beta}$$

$\therefore$ we get it

No.
Date. $\sum_{i=1}^{n} x_i \sim \text{Gamma}(n, \lambda)$ Denote it as $Y \sim \text{Gamma}(n, \lambda)$
$\sum_{i=1}^{n} X_i$

c) [redacted]

$$\therefore E[Y^{-1}] = E\left[\frac{1}{Y}\right] = \int_0^{+\infty} \frac{\frac{1}{y}\lambda^n \cdot y^{n-1}}{\Gamma(n)} e^{-y\lambda} dy = \int_0^{+\infty} \frac{y^{n-2}\lambda^n}{(n-1)\Gamma(n-1)} e^{-y\lambda} dy$$

$$= \frac{\lambda}{n-1} \boxed{\int_0^{+\infty} \frac{y^{n-2}\lambda^{n-1}}{\Gamma(n-1)} e^{-y\lambda} dy}$$

this part actually is another gamma distribution $\sim \text{Gamma}(n-1, \lambda)$

$\therefore$ it equals to 1

$\because$ the integral of PDF is 1

$$\therefore E[Y^{-1}] = \frac{\lambda}{n-1}$$

$$\therefore E\left[\frac{n-1}{n}\hat{\lambda}_{MLE}\right] = E\left[\frac{n-1}{n} \cdot \frac{n}{Y}\right] = n-1 \times E\left[\frac{1}{Y}\right]$$

$$= n-1 \times \frac{\lambda}{n-1} = \lambda$$

$\therefore$ we get it, so the $\frac{n-1}{n}\hat{\lambda}_{MLE}$ is unbiased.

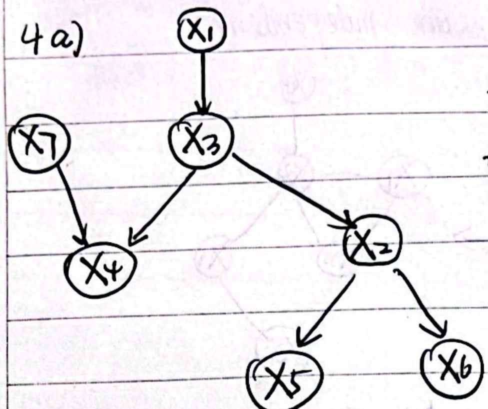
REMEMBER · MEMORY

王[illegible] 2021533135.

No.

Date. / /

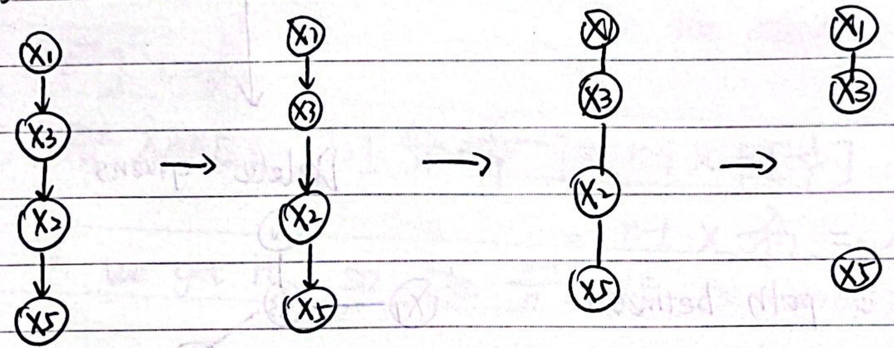
4 a)



the joint distribution of $X_1, \cdots X_7$, according the Bayes net is

$$P(X_1, X_2, \cdots, X_7) = P(X_1)P(X_3|X_1)P(X_7)P(X_4|X_3, X_7) \cdot P(X_2|X_3) \cdot P(X_5|X_2) \cdot P(X_6|X_2)$$

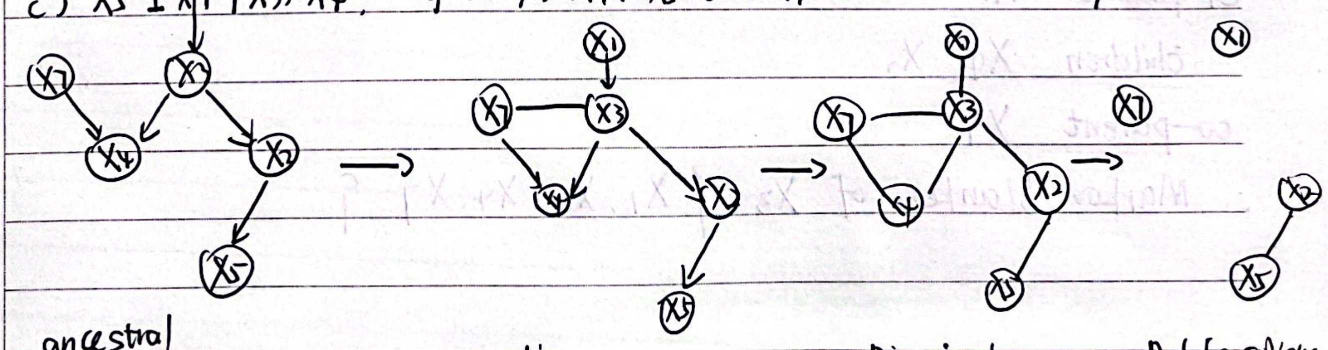
b) $X_1 \perp X_5 | X_2$. given X_2, X_1. X_5 whether are independent



ancestral graph   moralize   Disorient   Delete givens

$\therefore$ there is no path between X_1. X_5 nodes $\therefore X_1$. X_5 are independent, given X_2

c) $X_5 \perp X_7 | X_3, X_4$. given X_3. X_4. X_5 and X_7 whether are independent



ancestral graph   moralize   Disorient   Delete givens

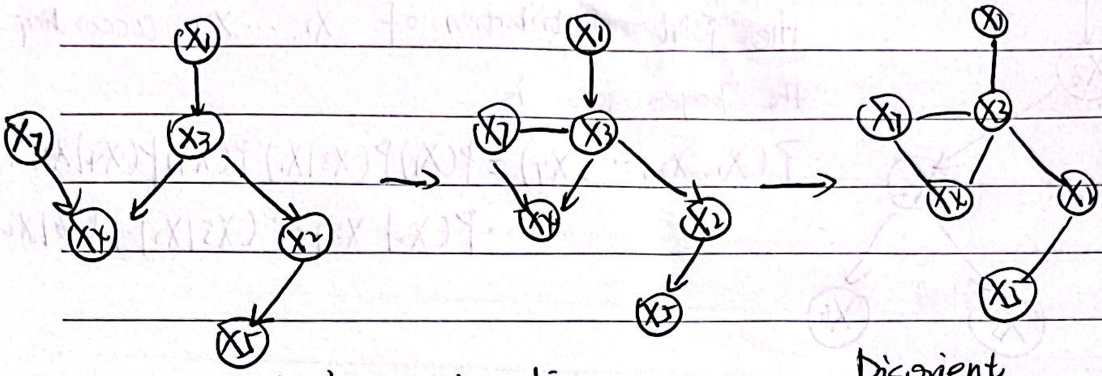
$\therefore$ there is no path between X_5. X_7 nodes

$\therefore X_5 \perp X_7 | X_3, X_4$

No.

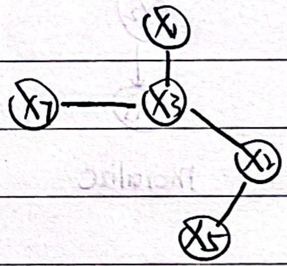
Date. / /

d) $X_5 \perp X_7 | X_4$ given X_4, X_5, X_7 whether are independent.



ancestral graph — moralize — Disorient

Delete givens



∴ there is a path between X_7 and X_5 nodes

∴ $X_5 \not\perp X_7 | X_4$,

X_5, X_7 are not conditional independent

e) parent X_1

children X_4 X_2

co-parent X_7

∴ Markov blanket of X_3: $\{X_1, X_2, X_4, X_7\}$