Machine Learning Written Assignment (Week-05) / 313652008 黃睿帆

1. $f(x) = \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$, where $x, \mu \in \mathbb{R}^k$, Σ is $k \times k$ matrix (positive-definite), $\underset{\sim}{\det = |\Sigma|}$.

$\boxed{\text{Want: } \int_{\mathbb{R}^k} f(x)\, dx = 1}$

sol: By linear algebra, we know that for positive-definite matrix, $\exists$ a unique positive definite matrix $A = |\Sigma|^{1/2}$ (A is square root of positive-definite matrix Σ) ← 矩陣平方根 $(|A|^2 = |\Sigma|)$

$\boxed{\begin{aligned} \Sigma &= AA^T \\ \Sigma^{-1} &= (A^T)^{-1} A^{-1} \end{aligned}}$

Change of variables: $z = A^{-1}(x-\mu) \Rightarrow x = \mu + Az$, where $dx = \underbrace{|A|}_{\text{(Jacobian)}} dz$.

$\boxed{z = A^{-1}(x-\mu)}$

Hence $f(x) = \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T (A^{-1})^T \cdot A^{-1}(x-\mu)} = \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(A^{-1}(x-\mu))^T \cdot (A^{-1}(x-\mu))} = \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(z^T z)}$

$\therefore \int_{\mathbb{R}^k} f(x)\, dx = \int_{\mathbb{R}^k} \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(z^T z)} \cdot |A|\, dz \underset{\left(\begin{aligned}|A| &= |\Sigma|^{1/2} \\ &= \sqrt{|\Sigma|}\end{aligned}\right)}{=} \int \dfrac{1}{\sqrt{(2\pi)^k}} e^{-\frac{1}{2}(z^T z)}\, dz$

(Noting that $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}_{n\times 1}$

$(a_1, \dots, a_n)\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = a_1^2 + a_2^2 + \dots + a_n^2$

i.e. $a^T a = \sum_{i=1}^{n} a_i^2$ (各分量平方和))

$= \dfrac{1}{\sqrt{(2\pi)^k}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}\sum_{i=1}^{k} z_i^2}\, dz \underset{\text{(積分可分離)}}{=} \left(\dfrac{1}{\sqrt{(2\pi)^k}}\right) \prod_{i=1}^{k} \int_{-\infty}^{\infty} e^{-\frac{z_i^2}{2}}\, dz_i = \dfrac{1}{\sqrt{(2\pi)^k}} \cdot (\sqrt{2\pi})^k = 1$

Gaussian integral: $\int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi}$

$(\therefore \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi})$

$\therefore \int_{\mathbb{R}^k} f(x)\, dx = 1$ #.

2. $A, B \in M_{n\times n}$ and x is $n\times 1$ vector

Observation, $A, B \in M_{3\times 3}$

$\left(\because \int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi}\right)$

$\left(\therefore \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}\right)$

2\. $A, B \in M_{n\times n}$ and x is $n\times 1$ vector

(a) $\because \operatorname{tr}(AB) = \sum_{i=1}^{n} (AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}$

For fixed index p, q, $\frac{\partial}{\partial A_{pq}} \operatorname{tr}(AB) = \frac{\partial}{\partial A_{pq}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}\right)$

Have term A_{pq} only when $i=p$ & $j=q$

$\boxed{即}$ The term $A_{pq}B_{qp}$.

$= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial A_{ij}}{\partial A_{pq}} B_{ji} = \sum_{i,j} \underbrace{\delta_{ip}\delta_{jq}}_{\boxed{\text{Kronecker Delta}}} B_{ji} = B_{qp}$.

$\therefore$ 對 A_{pq} 取偏導等於矩陣 B 在位置(q,p)的元素 B_{qp}.

$\Rightarrow \left(\frac{\partial \operatorname{tr}(AB)}{\partial A}\right)_{p,q} = B_{qp}$, for $p, q \in \{1, \dots, n\}$

$\therefore \frac{\partial \operatorname{tr}(AB)}{\partial A} = B^T$. #

Observation, $A, B \in M_{3\times 3}$

$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}_{3\times 3}$ $B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}_{3\times 3}$

$\Rightarrow AB = \begin{bmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32} & a_{11}b_{13}+a_{12}b_{23}+a_{13}b_{33} \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32} & a_{21}b_{13}+a_{22}b_{23}+a_{23}b_{33} \\ a_{31}b_{11}+a_{32}b_{21}+a_{33}b_{31} & a_{31}b_{12}+a_{32}b_{22}+a_{33}b_{32} & a_{31}b_{13}+a_{32}b_{23}+a_{33}b_{33} \end{bmatrix}$

$\operatorname{tr}(AB) = (a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31}) + (a_{21}b_{12}\ a_{22}b_{22}+a_{23}b_{32}) + (a_{31}b_{13}+a_{32}b_{23}+a_{33}b_{33})$

$\Rightarrow \frac{\partial(\operatorname{tr}AB)}{\partial A} = \begin{bmatrix} \frac{\partial \operatorname{tr}AB}{\partial a_{11}} & \frac{\partial \operatorname{tr}AB}{\partial a_{12}} & \frac{\partial \operatorname{tr}AB}{\partial a_{13}} \\ \vdots & \vdots & \vdots \\ \frac{}{\partial a_{21}} & \frac{}{\partial a_{22}} & \frac{}{\partial a_{23}} \\ \frac{}{\partial a_{31}} & \frac{}{\partial a_{32}} & \frac{}{\partial a_{33}} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{21} & b_{31} \\ b_{12} & b_{22} & b_{32} \\ b_{13} & b_{23} & b_{33} \end{bmatrix} = B^T$

① For scalar $s \in \mathbb{R}$, $s = \operatorname{tr}(s)$

② (Trace circulation) $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$

(Since by linear algebra, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.)

(b) Note that $(x^T)_{1\times n}(A)_{n\times n}(x)_{n\times 1} = (x^TAx)_{1\times 1}$ is a scalar, hence $\operatorname{tr}(x^TAx) \overset{①}{=} x^TAx$

then by ②, $\operatorname{tr}((x^TA)x) = \operatorname{tr}(x(x^TA)) = \operatorname{tr}(xx^TA)$. #

2. (c) $\boxed{\text{MLE}}$ (Max. Likelihood Estimation) of multi-variant Gaussian (normal) distribution :

$\left(\begin{array}{l} 2(a): \frac{\partial}{\partial A} \operatorname{trace}(AB) = B^T. \\ 2(b): x^T A x = tr(x x^T A). \end{array} \right)$

Suppose $X^{(1)}, X^{(2)}, \dots, X^{(m)}$ and $X^{(i)} \sim N_d(\mu, \Sigma)$, for all $i = 1, \dots, n$.

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

Then logistic likelihood function $\ell(\mu, \Sigma | x^{(i)}) = \ln \prod_{i=1}^{m} f_{x^{(i)}}(x^{(i)} | \mu, \Sigma) = \ln \prod_{i=1}^{m} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\left(\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)\right)}$

$$= \sum_{i=1}^{m} \left[-\frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma| - \frac{1}{2} (x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu) \right] \Rightarrow \ell(\mu, \Sigma) = -\frac{md}{2} \ln(2\pi) - \frac{m}{2} \ln|\Sigma| - \frac{1}{2} \sum_{i=1}^{m} (x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)$$

For fixed i, ($\mu = \mu^T$)

$(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)$

$= [(x^{(i)})^T - \mu][\Sigma^{-1} x^{(i)} - \Sigma^{-1} \mu]$

$= (x^{(i)})^T \Sigma^{-1} x^{(i)} - x^{(i)T} \Sigma^{-1} \mu - \mu \Sigma^{-1} x^{(i)} + \Sigma^{-1} \mu^2$

$\Rightarrow \frac{\partial}{\partial \mu} \Rightarrow -x^{(i)} \Sigma^{-1} - \Sigma^{-1} x^{(i)} + 2\mu \Sigma^{-1}$ 相等

$= 2\Sigma^{-1}(x^{(i)} - \mu)$

or Linear algebra : $\frac{\partial w^T A w}{\partial w} = 2A \cdot w$, if A is positive-definite (and w NOT depend on A)

To estimate $\hat{\mu} \Rightarrow \frac{\partial \ell(\mu, \Sigma)}{\partial \mu} = \left(-\frac{1}{2}\right)(-2) \Sigma^{-1} \sum_{i=1}^{m} (x^{(i)} - \mu) \overset{set}{=} 0$

$\Rightarrow \Sigma^{-1} \sum_{i=1}^{m} (x^{(i)} - \mu) = 0$, since Σ is positive-definite $\Rightarrow \sum_{i=1}^{m} (x^{(i)} - \mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} = \bar{X}$. #

and for covariance $\hat{\Sigma}$. $\left(-\frac{m}{2} \ln|\Sigma| = \frac{m}{2} \ln\left|\frac{1}{\Sigma}\right| \Rightarrow \frac{\partial}{\partial \Sigma} \frac{m}{2} \ln\left|\frac{1}{\Sigma}\right| = \frac{m}{2} \frac{1}{\left(\frac{1}{\Sigma}\right)} = \frac{m}{2} \Sigma \right)$

$$\Rightarrow \frac{\partial \ell(\mu, \Sigma)}{\partial \Sigma} = \left(+\frac{m}{2}\right) \Sigma - \frac{1}{2} \sum_{i=1}^{m} \frac{\partial [(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)]}{\partial \Sigma} = \frac{m}{2} \Sigma - \frac{1}{2} \sum_{i=1}^{m} (x^{(i)}-\mu)(x^{(i)}-\mu)^T \overset{set}{=} 0$$

$2(b): (x^{(i)}-\mu)^T \Sigma^{T} (x^{(i)}-\mu) = tr\left((x^{(i)}-\mu)(x^{(i)}-\mu)^T \Sigma\right)$ $\quad \Sigma^T = \Sigma = \Sigma^{-1}$

Then by $2(a)$, $\frac{\partial\, tr\left(\Sigma (x^{(i)}-\mu)(x^{(i)}-\mu)^T\right)}{\partial \Sigma} = (x^{(i)}-\mu)(x^{(i)}-\mu)^T$ $\quad \boxed{(AB)^T = B^T A^T}$

$$\Rightarrow \hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T \quad \#$$

Hence by MLE, multivariate Gaussian :

$$\begin{cases} \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \\ \hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T \end{cases} \quad \#.$$

◎ Unanswered Questions :

$$\Rightarrow \frac{\partial \ell(\mu,\Sigma)}{\partial \Sigma} = \left(+\frac{m}{2}\right)\Sigma - \frac{1}{2}\sum_{i=1}^{m} \frac{\partial\left[(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)\right]}{\partial \Sigma} = \frac{m}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{m} (x^{(i)}-\mu)(\ldots$$

$(\Sigma^T = \Sigma) \; \Sigma$

$2(b): (x^{(i)}-\mu)^T \Sigma^T (x^i - \mu) = tr\left((x^{(i)}-\mu)(x^{(i)}-\mu)^T \Sigma\right)$

Then by $2(a)$, $\frac{\partial\, tr\left(\Sigma (x^{(i)}-\mu)(x^{(i)}-\mu)^T\right)}{\partial \Sigma} = (x^{(i)}-\mu)(x^{(i)}-\mu)^T$

$\boxed{(AB)^T = B^T A^T}$

$$\Rightarrow \hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)}-\hat{\mu})(x^{(i)}-\hat{\mu})^T \quad \#$$

Hence by MLE, multivariate Gaussian :

$$\begin{cases} \hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \\ \hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)}-\hat{\mu})(x^{(i)}-\hat{\mu})^T \end{cases} \quad \#$$

◎ Unanswered Questions :

- In generative model, we assume, in the class, that $P(x|y=1)$, $P(x|y=0)$ are Gaussian & $P(y=1)$, $P(y=0)$ are Bernoulli, is it possible for other cases ? (Such as poisson, exponential, uniform distribution, etc.) and may we assume all components in $P(x|y)$ are from same distribution ? (such as $P(x|y=1)$, $P(y=1)$, $P(x|y=0)$, $P(y=0)$ are all Bernoulli ?)