ML AS5 111652042 顏友君

1. Given

$$f(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

where $x, \mu \in \mathbb{R}^k$, Σ is a k-by-k positive definite matrix and $|\Sigma|$ is its determinant. Show that $\int_{\mathbb{R}^k} f(x)\,dx = 1$.

① Σ^{-1} and $\Sigma^{\frac{1}{2}}$ existence :

Since Σ is a positive definite matrix iff Σ is symmetric and has all positive eigenvalues, we have Σ^{-1} exists, and $\Sigma = Q\Lambda Q^T$.

, where Q is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$

So, we can define $\Sigma^{\frac{1}{2}} = Q\Lambda^{\frac{1}{2}}Q$, where $\Lambda^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_k})$ s.t. $\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}} = \Sigma$.

② Change of variable

Let $y = \Sigma^{-\frac{1}{2}}(x-\mu)$.

Then, $x = \Sigma^{\frac{1}{2}}y + \mu \Rightarrow dx = |\Sigma^{\frac{1}{2}}|\,dy = |\Sigma|^{\frac{1}{2}}\,dy$.

Substitute $x - \mu = \Sigma^{\frac{1}{2}}y$, we have

$$\int_{\mathbb{R}^k} f(x)\,dx = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \cdot \int_{\mathbb{R}^k} e^{-\frac{1}{2}y^T\Sigma^{\frac{1}{2}}\cdot\Sigma^{-1}\cdot\Sigma^{\frac{1}{2}}y}\left(|\Sigma|^{\frac{1}{2}}dy\right) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \cdot \int_{\mathbb{R}^k} e^{-\frac{1}{2}y^Ty}|\Sigma|^{\frac{1}{2}}dy$$

$$= \frac{|\Sigma|^{\frac{1}{2}}}{\sqrt{(2\pi)^k|\Sigma|}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}y^Ty}dy = \frac{1}{(2\pi)^{\frac{k}{2}}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}y^Ty}dy \quad \text{— (i)}$$

③ Determine $\int_{\mathbb{R}^k} e^{-\frac{1}{2}y^Ty}dy$

Since $y^Ty = \sum_{i=1}^k y_i^2$, $e^{-\frac{1}{2}y^Ty} = e^{-\frac{1}{2}(\sum_{i=1}^k y_i^2)} = \prod_{i=1}^k e^{-\frac{1}{2}y_i^2}$.

Therefore, $\int_{\mathbb{R}^k} e^{-\frac{1}{2}y^Ty}dy = \int_{\mathbb{R}^k}\left(\prod_{i=1}^k e^{-\frac{1}{2}y_i^2}\right)dy = \prod_{i=1}^k\left(\int_{\mathbb{R}} e^{-\frac{1}{2}y_i^2}dy_i\right)$. — (ii)

<u>Claim that $\int_{\mathbb{R}} e^{-\frac{1}{2}y^2}dy = \sqrt{2\pi}$.</u>

Let $I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}dy$.

So, $I^2 = \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}dy\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}dx\,dy$.

Let $x = r\cos\theta$, $y = r\sin\theta$.

Then $dx\,dy = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} dr\,d\theta = r\,dr\,d\theta$

So, $I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2}r\,dr\,d\theta \overset{\text{let } u=\frac{r^2}{2}}{\underset{du=r\,dr}{=}} \int_0^{2\pi}\int_0^{\infty} e^{-u}du\,d\theta = 2\pi\left(-e^{-u}\big|_0^{\infty}\right) = 2\pi\cdot 1$.

Therefore, $I = \sqrt{2\pi} = \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}dy$. — (iii)

By (ii) and (iii), $\prod_{i=1}^k\left(\int_{\mathbb{R}} e^{-\frac{1}{2}y_i^2}dy_i\right) = k\cdot(\sqrt{2\pi}) = (2\pi)^{\frac{k}{2}}$. — (iv)

④ Finally, by (i) and (iv), $\int_{\mathbb{R}^k} f(x)\,dx = \frac{1}{(2\pi)^{\frac{k}{2}}}\cdot(2\pi)^{\frac{k}{2}} = 1$ □

2. Let A, B be n-by-n matrices and x be a n-by-1 vector.

(a) Show that $\frac{\partial}{\partial A}\text{trace}(AB) = B^T$.

(b) Show that $x^T A x = \text{trace}(x x^T A)$.

(c) Derive the maximum likelihood estimators for a multivariate Gaussian.

(p.s. 我寫完才發現我把(a)(b)的 n 都寫成 k了.
但應該不影響啦…)

(a) Given that A, B be k-by-k matrices, we have

$$\text{trace}(AB) = \text{trace}\left(\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k} \\ a_{2,1} & a_{2,2} & & \vdots \\ \vdots & & \ddots & \\ a_{k,1} & \cdots & & a_{k,k} \end{bmatrix}\begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,k} \\ b_{2,1} & b_{2,2} & & \vdots \\ \vdots & & \ddots & \\ b_{k,1} & \cdots & & b_{k,k} \end{bmatrix}\right)$$

$$= \text{trace}\left(\begin{bmatrix} \sum_{j=1}^{k} a_{1j} b_{j1} & & \\ & \sum_{j=1}^{k} a_{2j} b_{j2} & \\ & & \ddots \\ & & & \sum_{j=1}^{k} a_{kj} b_{jk} \end{bmatrix}\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}$$

doesn't matter

doesn't matter

Therefore, $\frac{\partial}{\partial A}\text{trace}(AB) = \frac{\partial}{\partial A}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right)$

$$= \begin{bmatrix} \frac{\partial}{\partial a_{11}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) & \frac{\partial}{\partial a_{12}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) & \cdots & \frac{\partial}{\partial a_{1k}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) \\ \frac{\partial}{\partial a_{21}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) & \frac{\partial}{\partial a_{22}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial}{\partial a_{k1}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) & & \cdots & \frac{\partial}{\partial a_{kk}}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} b_{ji}\right) \end{bmatrix} = \begin{bmatrix} b_{11} & b_{21} & \cdots & b_{k1} \\ b_{12} & b_{22} & & \vdots \\ \vdots & & \ddots & \\ b_{1k} & & \cdots & b_{kk} \end{bmatrix} = B^T \quad \square$$

(b) $$x^T A x = [x_1 \ x_2 \ \cdots \ x_k]\begin{bmatrix} a_{11} & \cdots & a_{1k} \\ a_{21} & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = \left[\sum_{i=1}^{k} x_i a_{i1} \ \ \sum_{i=1}^{k} x_i a_{i2} \ \cdots \ \sum_{i=1}^{k} x_i a_{ik}\right]\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = \sum_{j=1}^{k} x_j \sum_{i=1}^{k} x_i a_{ij} \text{ — (i)}$$

$$\text{trace}(x x^T A) = \text{trace}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}[x_1 \ \cdots \ x_k]\begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}\right) = \text{trace}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}\left[\sum_{i=1}^{k} x_i a_{i1} \ \ \sum_{i=1}^{k} x_i a_{i2} \ \cdots \ \sum_{i=1}^{k} x_i a_{ik}\right]\right)$$

$$= \text{trace}\left(\begin{bmatrix} x_1 \sum_{i=1}^{k} x_i a_{i1} & \cdots & x_1 \sum_{i=1}^{k} x_i a_{ik} \\ \vdots & x_2 \sum_{i=1}^{k} x_i a_{i2} & \\ x_k \sum_{i=1}^{k} x_i a_{i1} & \cdots & x_k \sum_{i=1}^{k} x_i a_{ik} \end{bmatrix}\right) = \sum_{j=1}^{k} x_j \sum_{i=1}^{k} x_i a_{ij} \text{ — (ii)}$$

Comparing (i) and (ii), $x^T A x = \sum_{j=1}^{k} x_j \sum_{i=1}^{k} x_i a_{ij} = \text{trace}(x x^T A)$ $\square$

(c) Let $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ be independent samples s.t. $x_i \sim N(\mu, \Sigma)$, for $i = 1, \ldots, N$, where $\mu \in \mathbb{R}^n$ is the mean vector and $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix.

The pdf of the multivariate Gaussian is

$$p(x_i \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}$$

Assuming the samples are independent, the likelihood function is

$$L(\mu, \Sigma) = \prod_{i=1}^{N} p(x_i \mid \mu, \Sigma) = \left(\frac{1}{((2\pi)^n |\Sigma|)^{\frac{1}{2}}}\right)^N e^{-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}$$

Therefore, $\ell(\mu, \Sigma) = -\frac{nN}{2}\log(2\pi) - \frac{N}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$.

To find MLE for μ, take $\frac{\partial \ell}{\partial \mu} = 0$.

We have, $\frac{\partial \ell}{\partial \mu} = -\frac{1}{2} \cdot -2 \cdot \Sigma^{-1} \sum_{i=1}^{N}(x_i - \mu) = \Sigma^{-1} \sum_{i=1}^{N}(x_i - \mu) = \Sigma^{-1}\left(\left(\sum_{i=1}^{N} x_i\right) - N\mu\right) = 0$

So, $\underline{\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i}$ #

To find MLE for the covariance, take $\frac{\partial \ell}{\partial \Sigma} = 0$

Substitute $\mu = \hat{\mu}$ (as a constant) into $\ell(\mu, \Sigma)$,

$$\ell(\hat{\mu}, \Sigma) = -\frac{nN}{2}\log(2\pi) - \frac{N}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\hat{\mu})^T\Sigma^{-1}(x_i-\hat{\mu})$$

$$= -\frac{nN}{2}\log(2\pi) - \frac{N}{2}\log(|\Sigma|) - \frac{1}{2}\operatorname{trace}(\Sigma^{-1}S), \text{ where } S = \sum_{i=1}^{N}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$$

Since $\frac{\partial}{\partial \Sigma}\log|\Sigma| = \Sigma^{-1}$ and $\frac{\partial}{\partial \Sigma}\operatorname{trace}(\Sigma^{-1}S) = -\Sigma^{-1}S\Sigma^{-1}$. (by GPT)

we have $\frac{\partial \ell}{\partial \Sigma} = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}S\Sigma^{-1} = 0$

So, $\Sigma^{-1}S\Sigma^{-1}\Sigma = N\Sigma^{-1}\Sigma \Rightarrow \Sigma^{-1}S = NI \Rightarrow \underline{\hat{\Sigma}_{MLE} = \frac{1}{N}\cdot S = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})(x_i-\hat{\mu})^T}$ #

3. Unanswered Question:

在課堂中提到的LDA主要是用來做分類(Classification),

但LDA也可以用來做降維(dimension reduction),如果LDA用來做降維的話,還需要有標籤(labels)嗎?若需要的話,是否也需要假設資料為Gaussian distribution?