Week 5 written assignment

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

我們的目標是計算以下積分：

$$I = \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\, dx$$

將常數項提出積分外：

$$I = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\, dx$$

為了簡化指數項，我們使用變數變換。由於 Σ 是一個對稱正定矩陣，它可以被分解為 $\Sigma = CC^T$，其中 C 是一個可逆矩陣。因此，$\Sigma^{-1} = (C^T)^{-1}C^{-1}$。

我們定義新的變數 $y \in \mathbb{R}^k$：

$$y = C^{-1}(x - \mu)$$

由此可得：

$$x = Cy + \mu$$

接下來，我們需要計算這個變換的 Jacobian 行列式。x 對 y 的導數矩陣為：

$$\frac{\partial x}{\partial y} = C$$

Jacobian 行列式為 $|J| = |\det(C)|$。因為 $\det(\Sigma) = \det(CC^T) = \det(C)\det(C^T) = (\det(C))^2$，所以 $|\det(C)| = \sqrt{|\Sigma|}$。

將積分式中的 x 替換為 y。首先看指數項：

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = (Cy)^T(C^T)^{-1}C^{-1}(Cy) = y^TC^T(C^T)^{-1}C^{-1}Cy = y^TIy = y^Ty = \sum_{i=1}^{k} y_i^2$$

積分的微分元素 dx 變為 $|J|\,dy = \sqrt{|\Sigma|}\,dy$。將這些代回原積分式：

$$I = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}\sum_{i=1}^k y_i^2} \sqrt{|\Sigma|}\,dy$$

$\sqrt{|\Sigma|}$ 項可以消去：

$$I = \frac{1}{\sqrt{(2\pi)^k}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}(y_1^2+y_2^2+\cdots+y_k^2)}\,dy_1\,dy_2 \ldots dy_k$$

這個多維積分可以分解為 k 個獨立的一維積分的乘積：

$$I = \frac{1}{(2\pi)^{k/2}} \left(\int_{-\infty}^{\infty} e^{-\frac{y_1^2}{2}}\,dy_1\right) \left(\int_{-\infty}^{\infty} e^{-\frac{y_2^2}{2}}\,dy_2\right) \ldots \left(\int_{-\infty}^{\infty} e^{-\frac{y_k^2}{2}}\,dy_k\right)$$

我們知道標準一維高斯積分的結果是：

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

因此，我們的積分 I 變為：

$$I = \frac{1}{(2\pi)^{k/2}}(\sqrt{2\pi})^k = \frac{(2\pi)^{k/2}}{(2\pi)^{k/2}} = 1$$

2

2.1 證明 $\frac{\partial}{\partial A}\textbf{trace}(AB) = B^T$

設 A 和 B 均為 $n \times n$ 矩陣。我們將 trace 寫成元素求和的形式：

$$\text{trace}(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}$$

我們要求這個純量對矩陣 A 的導數。其結果是一個矩陣，該矩陣的 (k,l) 位置的元素是 $\frac{\partial}{\partial A_{kl}}\text{trace}(AB)$。

$$\frac{\partial}{\partial A_{kl}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial A_{ij}}{\partial A_{kl}} B_{ji}$$

根據偏微分的定義，$\frac{\partial A_{ij}}{\partial A_{kl}}$ 只有在 $i = k$ 且 $j = l$ 時為 1，否則為 0。因此，上式變為：

$$\frac{\partial}{\partial A_{kl}}\text{trace}(AB) = B_{lk}$$

這剛好是矩陣 B^T 在 (k,l) 位置的元素。因此，可以得出結論：

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

2.2 證明 $x^T A x = \textbf{trace}(x x^T A)$

設 x 為 $n \times 1$ 向量，A 為 $n \times n$ 矩陣。$x^T A x$ 的計算結果是一個 1×1 的矩陣，即一個純量。對於任何純量 c，都有 $c = \text{trace}(c)$。因此：

$$x^T A x = \text{trace}(x^T A x)$$

因為 $\text{trace}(AB) = \text{trace}(BA)$ ，所以

$$\text{trace}(x^T A x) = \text{trace}(A x x^T) = \text{trace}((x x^T) A)$$

因此得到：

$$x^T A x = \text{trace}(x x^T A)$$

2.3 推導多變量高斯分佈的最大概似估計

假設我們有一組從多變量高斯分佈 $N(\mu, \Sigma)$ 中獨立同分佈 (i.i.d.) 抽樣得到的數據點 $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$。

1. 概似函數是所有數據點機率的連乘積：

$$L(\mu, \Sigma) = \prod_{i=1}^{m} p(x^{(i)}; \mu, \Sigma) = \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)}$$

取 log 得到對數概似函數 $\ell(\mu, \Sigma) = \ln L(\mu, \Sigma)$：

$$\ell(\mu, \Sigma) = -\frac{mk}{2} \ln(2\pi) - \frac{m}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{m} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)$$

2. 為了找到使 ℓ 最大化的 μ，我們計算 ℓ 對 μ 的偏微分並使其為零。

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

令 $\frac{\partial \ell(\mu, \Sigma)}{\partial \mu} = 0 \Rightarrow$

$$\Sigma^{-1} \sum_{i=1}^{m} (x^{(i)} - \mu) = 0 \implies \sum_{i=1}^{m} (x^{(i)} - \mu) = 0 \implies \sum_{i=1}^{m} x^{(i)} - m\mu = 0$$

解得 μ 的最大概似估計 $\hat{\mu}$：

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

3. 為了找到使 ℓ 最大化的 Σ，我們計算 ℓ 對 Σ^{-1} 的導數。令 $S = \Sigma^{-1}$，則 $|\Sigma| = |S|^{-1}$。

$$\ell(\mu, S) = C + \frac{m}{2} \ln |S| - \frac{1}{2} \sum_{i=1}^{m} \text{trace}\left((x^{(i)} - \mu)(x^{(i)} - \mu)^T S \right)$$

對 S 求導，並利用求導法則 $\frac{\partial}{\partial A}\ln|A| = (A^{-1})^T$ 和 $\frac{\partial}{\partial A}\text{trace}(BA) = B^T$：

$$\frac{\partial \ell}{\partial S} = \frac{m}{2}(S^{-1})^T - \frac{1}{2}\sum_{i=1}^{m}\left((x^{(i)} - \mu)(x^{(i)} - \mu)^T\right)^T$$

因為 S 和 $(x^{(i)} - \mu)(x^{(i)} - \mu)^T$ 都是對稱矩陣，轉置等於自身：

$$\frac{\partial \ell}{\partial S} = \frac{m}{2}S^{-1} - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

令導數為零，解得 S^{-1}：

$$S^{-1} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

由於 $S = \Sigma^{-1}$，所以 $S^{-1} = \Sigma$。將上面求得的 $\hat{\mu}$ 代入，得到 Σ 的最大概似估計 $\hat{\Sigma}$：

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

3 Unanswered Questions

上課介紹了生成模型 GDA (包含 LDA 和 QDA) 和 Logistic regression。在實際應用中，當我們面對一個分類問題時，應如何在這兩類模型之間進行選擇？它們各自的優勢、劣勢以及適用的數據情境是什麼？例如，在數據量較小或特徵之間相關性很強的情況下，哪種模型可能表現更佳？