(1) 111652026 劉馥瑞

Let $y = \Sigma^{-\frac{1}{2}}(x-\mu)$, where $\Sigma^{-\frac{1}{2}}$ is the symmetric square root of Σ^{-1}.

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

So, $(x-\mu)^T \Sigma^{-1}(x-\mu) = (x-\mu)^T (\Sigma^{-\frac{1}{2}})^T \cdot \Sigma^{-\frac{1}{2}}(x-\mu) = y^T y = \|y\|^2$.

Substitute into $f(x)$: $f(x) = \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp(-\frac{1}{2}\|y\|^2)$

Integrating both sides : $\int_{\mathbb{R}^k} f(x)dx = \int_{\mathbb{R}^k} \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp(-\frac{1}{2}\|y\|^2)\, |\Sigma|^{\frac{1}{2}} dy$

$= \int_{\mathbb{R}^k} \dfrac{1}{(2\pi)^{\frac{k}{2}}} e^{-\frac{1}{2}\|y\|^2} dy$, and this form is the integral of the standard multivariate normal density, hence, $\int_{\mathbb{R}^k} f(x)dx = 1$. □

2.

(a)

$\text{trace}(AB) = \sum_{i,j} A_{ij} B_{ji}$.

Then, $\frac{\partial}{\partial A_{ij}} \text{trace}(AB) = B_{ji}$.

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

(b)

Note that $x^T A x$ is a scalar, and $\text{trace}(x^T A x) = \text{trace}((x^T A)x)$.

Using that $\text{trace}(uv^T) = v^T u$, so $x^T A x = \text{trace}((x^T A)x) = \text{trace}(x x^T A)$.

(c)

Suppose we have i.i.d. samples $x_1, \ldots, x_m \sim N(\mu, \Sigma)$

The likelihood : $L(\mu, \Sigma) = \prod_{i=1}^{m} \dfrac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right)$.

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

Differentiate w.r.t. μ.

$$\frac{\partial l}{\partial \mu} = \Sigma^{-1} \sum_{i=1}^{m} (x_i - \mu).$$

set $= 0$:

$$\sum_{i=1}^{m} (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i .$$

Differentiate w.r.t. Σ.

$\frac{\partial}{\partial \Sigma} \left(-\frac{m}{2} \ln |\Sigma| - \frac{1}{2} \Sigma (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right) = 0$ by matrix calculus results.

That yields: $\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$

3.

Why does LDA use class means and a pooled covariance matrix ?