

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

<pf>

1° For $\Sigma_{k \times k}$ is symmetric and positive definite, there exists an invertible matrix A s.t. $\Sigma = AA^T$, moreover $|\Sigma| = (\det A)^2$

Now, we define the new variable $y = A^{-1}(x-\mu) \Rightarrow x = \mu + Ay$

Moreover, we have the Jacobian determinant of this transformation is

$$dx = |\det A| dy = \sqrt{|\Sigma|} dy$$

2° Substituting $x = \mu + Ay$ into our exponent:

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = y^T A^T (AA^T)^{-1} Ay = y^T y = \|y\|^2$$

$$3^\circ \int_{\mathbb{R}^k} f(x) dx = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) dx$$

By 2°

$$= \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}\|y\|^2\right) |\det A| dy$$

$$= \frac{|\det A|}{\sqrt{(2\pi)^k |\Sigma|}} \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}\|y\|^2\right) dy$$

$$\text{where, } \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}\|y\|^2\right) dy = \prod_{j=1}^k \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_j^2} dy_j = (\sqrt{2\pi})^k = (2\pi)^{\frac{k}{2}}$$

$$\text{Therefore, } \int_{\mathbb{R}^k} f(x) dx = \frac{|\det A|}{\sqrt{(2\pi)^k |\Sigma|}} (2\pi)^{\frac{k}{2}} = 1 \quad \blacksquare$$

$$\begin{aligned} \times \int_{-\infty}^{\infty} e^{-ax^2} dx &= \sqrt{\frac{\pi}{a}} \\ \text{Then, } \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_j^2} dy_j &= \sqrt{2\pi} \end{aligned}$$

By Calculus 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.

(a)

<pf> 1° Suppose that $A = [a_{ij}]$, $B = [b_{ij}]$, then $(AB)_{ii} = \sum_{k=1}^n a_{ik} b_{ki}$. Therefore, $\text{tr}(AB) = \sum_{i=1}^n \sum_{k=1}^n a_{ik} b_{ki}$

2° We take the partial derivative for a_{mn} .

We can see that the only term which depend on a_{mn} is when $i=m$; $k=n$.

$$\text{Thus, the derivative} = \frac{\partial}{\partial a_{mn}} \text{tr}(AB) = b_{mn}$$

1° The gradient $\frac{\partial}{\partial A} \text{tr}(AB)$ is the matrix with (m,n) -entry is exactly this derivative:

$$\text{i.e. } \left\{ \frac{\partial}{\partial A} \text{tr}(AB) \right\}_{mn} = \frac{\partial}{\partial a_{mn}} \text{tr}(AB) = b_{mn}$$

However, the matrix of B^T has entries $(B^T)_{mn} = b_{nm}$.

Therefore, we prove that $\frac{\partial}{\partial A} \text{tr}(AB) = B^T$ \square

2° method 2: "Differential + Frobenius inner product"

1° Let $f(A) = \text{tr}(AB)$, where B is a constant

Then, the differential of f is: $df = d \text{tr}(AB) = \text{tr}(dAB) = \text{tr}(dA B)$

(for B doesn't depend on A)

2° Using the cyclic property of the trace:

Using $\text{tr}(dAB) = \text{tr}(B dA) = \text{tr}((B^T)^T dA) = \langle B^T, dA \rangle_F$, where $\langle X, Y \rangle_F = \text{tr}(X^T Y)$ Frobenius inner product.

3° By the definition of the matrix gradient: $df = \langle \nabla_A f, dA \rangle_F$

However, we also know that $df = \langle B^T, dA \rangle_F$

Since this must hold for all variation dA , we conclude: $\nabla_A f = \frac{\partial}{\partial A} \text{tr}(AB) = B^T$ \square

(b) 1° $X^T A X$ is a scalar (i.e. $X^T A X$ is a 1×1 matrix)

$\forall S$ is scalar, we have $S = \text{tr}(S)$

For the trace of a 1×1 matrix is just the value itself (i.e. $X^T A X = \text{tr}(X^T A X)$)

2° Using the cyclic property of trace:

(for example: $\text{tr}(PQR) = \text{tr}(QRP) = \text{tr}(RPQ)$), here $P = X^T_{1 \times n}$; $Q = A_{n \times n}$; $R = X_{n \times 1}$

Then, we have $\text{tr}(X^T A X) = \text{tr}(A X X^T) = \text{tr}(X X^T A)$

3° The matrix $X X^T$ is an $n \times n$ outer product, so the final expression $\text{tr}(X X^T A)$ is exactly the same with the original scalar $X^T A X$

Therefore, $X^T A X = \text{tr}(X X^T A)$ \square

(c) 1° Suppose that we have m independent samples:

$\{x_1, x_2, \dots, x_m\} \in \mathbb{R}^k$, where $x_i \sim \mathcal{N}(\mu, \Sigma)$, with unknown mean vector $\mu \in \mathbb{R}^k$ and the co-variance matrix $\Sigma \in \mathbb{R}^{k \times k}$ (symmetric positive definite).

2° The probability density function for one sample:

$$p(x_i | \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right],$$

Moreover, the likelihood function for the full data-set: $\mathcal{L}(\mu, \Sigma) = \prod_{i=1}^m p(x_i | \mu, \Sigma)$

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

* Define the scatter matrix $S(\mu) = \sum_{i=1}^m (x_i - \mu)(x_i - \mu)^T$

Using the result of (b), then we have $\sum_{i=1}^m (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) = \text{tr}(\Sigma^{-1} S(\mu))$

Therefore, $l(\mu, \Sigma) = -\frac{m}{2} \ln |\Sigma| - \frac{1}{2} \text{tr}(\Sigma^{-1} S(\mu))$

$$3^{\circ} \frac{\partial l}{\partial \mu} = -\frac{1}{2} \left(\frac{\partial}{\partial \mu} \text{tr}(\Sigma^{-1} S(\mu)) \right) = \Sigma^{-1} \left(\sum_{i=1}^m x_i - m\mu \right) \leftarrow \text{Expanding } S(\mu) \text{ and the differentiating.}$$

Suppose that $\Sigma^{-1} \left(\sum_{i=1}^m x_i - m\mu \right) = 0$, then $\sum_{i=1}^m x_i - m\mu = 0$, we have $\hat{\mu} = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$

$$4^{\circ} \frac{\partial}{\partial \Sigma} \ln |\Sigma| = \Sigma^{-1} ; \frac{\partial}{\partial \Sigma} \text{tr}(\Sigma^{-1} S) = -\Sigma^{-T} S \Sigma^{-T}$$

$$\text{then, } \frac{\partial l}{\partial \Sigma} = -\frac{m}{2} \Sigma^{-T} + \frac{1}{2} \Sigma^{-T} S(\hat{\mu}) \Sigma^{-T}$$

Suppose that $-\frac{m}{2} \Sigma^{-T} + \frac{1}{2} \Sigma^{-T} S(\hat{\mu}) \Sigma^{-T} = 0$, moreover, multiply on the left and right by Σ^T

we can get $S(\hat{\mu}) = m\Sigma$

$$\text{Thus, } \hat{\Sigma} = \frac{1}{m} S(\hat{\mu}) = \frac{1}{m} \sum_{i=1}^m (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

3. Unanswered Questions

There are unanswered questions from the lecture, and there are likely more questions we haven't covered.

Q: If there is a multiclass classification, how do we find the boundary?