HW5

112652011 廖晨鈞

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

Solution

Solve the integral

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

First, let $y = x - \mu \in \mathbb{R}^k$, $dx = dy$.

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

Since Σ is a positive definite matrix, so is Σ^{-1}. By the Spectral theorem, we have

$$\Sigma^{-1} = PDP^T$$

where P is a orthogonal matrix, D is a diagonal matrix with entries $\lambda_1, \lambda_2, \ldots, \lambda_k$.

Let $y = Pz$, $dy = |\det(P)| dz = dz$.

So,

$$y^T \Sigma^{-1} y = (Pz)^T (PDP^T)(Pz) = z^T P^T PDP^T Pz$$

Since $P^T P = I$,

$$y^T \Sigma^{-1} y = z^T D z = \sum_{i=1}^{k} \lambda_i z_i^2$$

So,

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

Moreover, we need to write $|\Sigma|$ in the form of λ_i

$$|\Sigma^{-1}| = |PDP^T| = |P||D||P^T| = |D| = \prod_{i=1}^{k} \lambda_i$$

So,

$$|\Sigma| = \frac{1}{|\Sigma^{-1}|} = \frac{1}{\prod_{i=1}^{k} \lambda_i}$$

This integral become

$$\begin{aligned}
I &= \int_{\mathbb{R}^k} \left(\prod_{i=1}^{k} \sqrt{\frac{\lambda_i}{2\pi}} \right) e^{-\frac{1}{2}\sum_{i=1}^{k} \lambda_i z_i^2} \, dz \\
&= \left(\prod_{i=1}^{k} \sqrt{\frac{\lambda_i}{2\pi}} \right) \int_{\mathbb{R}^k} e^{-\frac{1}{2}\sum_{i=1}^{k} \lambda_i z_i^2} \, dz \\
&= \left(\prod_{i=1}^{k} \sqrt{\frac{\lambda_i}{2\pi}} \right) \int_{\mathbb{R}^k} e^{-\frac{1}{2}\lambda_1 z_1^2} e^{-\frac{1}{2}\lambda_2 z_2^2} \dots e^{-\frac{1}{2}\lambda_k z_k^2} \, dz_1 dz_2 \dots dz_k \\
&= \left(\prod_{i=1}^{k} \sqrt{\frac{\lambda_i}{2\pi}} \right) \int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_1 z_1^2} dz_1 \int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_2 z_2^2} dz_2 \dots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_k z_k^2} dz_k \\
&= \left(\prod_{i=1}^{k} \sqrt{\frac{\lambda_i}{2\pi}} \right) \prod_{i=1}^{k} \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}\lambda_i z_i^2} dz_i \right)
\end{aligned}$$

We use the 1-D Gaussian integral: $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$.
Here, $a = \frac{\lambda_i}{2}$, so

$$\int_{-\infty}^{\infty} e^{-\frac{\lambda_i}{2} z_i^2} \, dz_i = \sqrt{\frac{\pi}{\lambda_i/2}} = \sqrt{\frac{2\pi}{\lambda_i}}$$

Substitute back to I, we get

$$I = \prod_{i=1}^{k} \left(\sqrt{\frac{\lambda_i}{2\pi}} \cdot \sqrt{\frac{2\pi}{\lambda_i}} \right) = \prod_{i=1}^{k} 1 = 1$$

Problem 2

Let A, B be n-by-n matrices and x be a n-by-1 vector.
(a) Show that $\frac{\partial}{\partial A}\operatorname{trace}(AB) = B^T$.
(b) Show that $x^T Ax = \operatorname{trace}(xx^T A)$.
(c) Derive the maximum likelihood estimators for a multivariate Gaussian.

Solution

(a)

Write the trace in element form:

$$\operatorname{trace}(AB) = \sum_i \sum_j A_{ij} B_{ji}$$

Now, we calculate the partial derivative of the trace w.r.t. an arbitrary element A_{kl} of the matrix A

$$\frac{\partial}{\partial A_{kl}}\operatorname{trace}(AB) = \frac{\partial}{\partial A_{kl}}\left(\sum_i \sum_j A_{ij} B_{ji}\right)$$

In this double summation, The derivative is non-zero only for the term where $i = k$ and $j = l$. Thus,

$$\frac{\partial}{\partial A_{kl}}\left(\sum_i \sum_j A_{ij} B_{ji}\right) = \frac{\partial}{\partial A_{jk}}(A_{jk} B_{kj}) = B_{lk}$$

The resulting matrix has B_{lk} at the (k, l) position, which is the definition of B^T. So,

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

(b)

The term $x^T Ax$ is a scalar (a 1×1 matrix). So,

$$x^T Ax = \operatorname{trace}(x^T Ax)$$

The trace operator has a cyclic property: $\operatorname{trace}(ABC) = \operatorname{trace}(CAB)$. So,
$$\operatorname{trace}(x^T Ax) = \operatorname{trace}(xx^T A)$$

(c)

Given m i.i.d. samples $\{x^{(1)}, \ldots, x^{(m)}\}$, the log-likelihood function is:

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

1. Estimator for the mean μ:

Take the gradient of ℓ with respect to μ and set it to zero:

$$\nabla_\mu \ell = \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu) = \sum_{i=1}^{m}\Sigma^{-1}(x^{(i)}-\mu) = 0$$

Solving for μ gives the maximum likelihood estimator $\hat{\mu}$:

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

The MLE for the mean is the sample mean.

2. Estimator for the covariance Σ:

Rewrite the quadratic term using (b). This makes differentiation simpler.

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \text{trace}\left(\Sigma^{-1}(x-\mu)(x-\mu)^T\right)$$

The log-likelihood becomes

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

Differentiate with respect to Σ. We use two matrix derivative rules

- $\frac{\partial}{\partial A}\ln|A| = (A^{-1})^T$
- $\frac{\partial}{\partial A}\text{trace}(A^{-1}B) = -(A^{-1}BA^{-1})^T$

Applying these rules (and Σ is symmetric, so $\Sigma^T = \Sigma$), we get

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

Set the derivative to zero

$$-\frac{m}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left(\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T\right)\Sigma^{-1} = 0$$

Multiply from the left by 2Σ

$$-mI + \left(\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T\right)\Sigma^{-1} = 0$$

$$mI = \left(\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T\right)\Sigma^{-1}$$

Multiply from the right by $\frac{1}{m}\Sigma$

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

Substitute the estimator $\hat{\mu}$ to get the final result

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

The MLE for the covariance is the sample covariance matrix.

Unanswered Questions

What is the geometric meaning of the assumption that simplifies GDA to LDA (i.e., $\Sigma_0 = \Sigma_1 = \Sigma_0$)?