Classification Rule Derivation

April 29, 2020

0.0.1 a) Classification rule devivation

Starting with some definitions:

- π_i is the prior probability that an individual selected at random belongs to population i.
- $C(i|j)$ is the cost of incorrectly allocating an individual to population i, when they really belong to population j.

Bayes rule:

Allocate population 1 if:

$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}$,

otherwise allocate population 2.

In the multivariate normal case (with p variables), where the observation vectors $\mathbf{X}_i \sim \mathrm{MVN}_p(\mu_\mathbf{i}, \Sigma)$, then our two probability density functions for population 1 and 2 are:

$$f_1(\mathbf{x}) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{1})\right\},$$

and

$$f_2(\mathbf{x}) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{2})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{2})\right\},$$

respectively.

We can therefore calculate the likelihood ratio as:

$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{1}) + \frac{1}{2}(\mathbf{x} - \mu_\mathbf{2})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{2})\right\}$,

where the $\frac{1}{|2\pi\Sigma|^{1/2}}$ terms in each pdf cancel out.

We can then calculate the log-likelihood by taking natural logs of both sides:

$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{1}) + \frac{1}{2}(\mathbf{x} - \mu_\mathbf{2})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{2})$

and factorising the $\frac{1}{2}$:

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = -\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu_1})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) - (\mathbf{x}-\boldsymbol{\mu_2})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_2})\right).$$

We can expand out the $(\mathbf{x}-\boldsymbol{\mu_1})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) - (\mathbf{x}-\boldsymbol{\mu_2})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_2})$ terms:

Firstly $(\mathbf{x}-\boldsymbol{\mu_1})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_1})$

$$= (\mathbf{x}^T\Sigma^{-1} - \boldsymbol{\mu_1}^T\Sigma^{-1})(\mathbf{x}-\boldsymbol{\mu_1})$$
$$= \mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_1}^T\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1}$$

Each of these four components is a scalar, and the transpose of a scalar is the scalar itself, i.e. $a^T = a$, therefore:

$$\boldsymbol{\mu_1}^T\Sigma^{-1}\mathbf{x} = \left(\boldsymbol{\mu_1}^T\Sigma^{-1}\mathbf{x}\right)^T = \mathbf{x}^T(\Sigma^{-1})^T\boldsymbol{\mu_1} = \mathbf{x}^T(\Sigma^T)^{-1}\boldsymbol{\mu_1} = \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu_1},$$

since Σ is a symmetric matrix, $\Sigma = \Sigma^T$.

We can therefore write:

$$(\mathbf{x}-\boldsymbol{\mu_1})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) = \mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu_1} + \boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1}$$

and

$$(\mathbf{x}-\boldsymbol{\mu_2})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_2}) = \mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu_2} + \boldsymbol{\mu_2}^T\Sigma^{-1}\boldsymbol{\mu_2}.$$

Thus we can expand out:

$$(\mathbf{x}-\boldsymbol{\mu_1})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) - (\mathbf{x}-\boldsymbol{\mu_2})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_2})$$

to

$$-2\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu_1} + \boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1} + 2\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu_2} - \boldsymbol{\mu_2}^T\Sigma^{-1}\boldsymbol{\mu_2},$$

where we can factor out the $-2\mathbf{x}^T\Sigma^{-1}$ terms:

$$-2\mathbf{x}^T\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}) + \boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2}^T\Sigma^{-1}\boldsymbol{\mu_2}.$$

The $\boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2}^T\Sigma^{-1}\boldsymbol{\mu_2}$ terms follow the difference of two squares, and can be written as $(\boldsymbol{\mu_1}+\boldsymbol{\mu_2})^T\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$.

Our log-likelihood can therefore be written:

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = -\frac{1}{2}\left(-2\mathbf{x}^T\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}) + (\boldsymbol{\mu_1}+\boldsymbol{\mu_2})^T\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})\right)$$
$$= \mathbf{x}^T\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}) - \frac{1}{2}(\boldsymbol{\mu_1}+\boldsymbol{\mu_2})^T\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$$

The common factor in both terms is $\Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$. Let $\mathbf{L} = \Sigma^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$ such that our log-likelihood takes the form:

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = \mathbf{x}^T\mathbf{L} - \frac{1}{2}(\boldsymbol{\mu_1}+\boldsymbol{\mu_2})^T\mathbf{L}$$

Note again that each of the two above terms is a scalar ($\mathbf{L}$ is a $p \times 1$ vector), so we can shuffle the order of the terms by taking the transpose of each:

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = \mathbf{L}^T\mathbf{x} - \frac{1}{2}\mathbf{L}^T(\mu_{\mathbf{1}} + \mu_{\mathbf{2}}).$$

Recall that as:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)},$$

the log-likelihood relationship must be:

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) \geq \ln\left(\frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}\right),$$

therefore:

$$\mathbf{L}^T\mathbf{x} - \frac{1}{2}\mathbf{L}^T(\mu_{\mathbf{1}} + \mu_{\mathbf{2}}) \geq \ln\left(\frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}\right).$$

Our allocation rule is therefore:

Allocate to population 1 if:

$$\mathbf{L}^T\mathbf{x} - \frac{1}{2}\mathbf{L}^T(\mu_{\mathbf{1}} + \mu_{\mathbf{2}}) \geq \ln\left(\frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}\right),$$

otherwise allocate to population 2.
