Multivariate Data Analysis. Home Assignment 3

Xijia Liu

Part I: Computer exercises. In the next series of tasks, we understand the linear discriminant analysis (LDA), Fisher's projection, and the 'lda()' function.

Task 1: Generate $n_1 = 150$ and $n_2 = 150$ observations from bivariate normal distributions with $\boldsymbol{\mu}_1 = (0, 0)'$ and $\boldsymbol{\mu}_2 = (-3, 2)'$ and identical covariance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$, respectively. Assign label "1" to the first 150 observation and "2" to the rest. Visualize the simulated data with different colours in a figure.

Task 2: Assuming the two populations have an equal proportion, we fit an LDA model using the simulated observations. Write down the mathematical expression of the decision boundary of your LDA model, then calculate the maximum likelihood estimations (MLE) of all the unknown parameters, and add the decision boundary to the plot in Task 1. **Hint**: you can use 'abline' function to draw the decision boundary. **Note**: 'lda()' is not allowed to use here.

Task 3: Apply the 'lda()' function in 'MASS' package on your data set to fid the LDA model. Find the decision boundary given the outputs and compare it with the decision boundary in Task 2. Draw a conclusion based on Tasks 2 and 3. **Hint**: in the outputs, you can find one slot which is called "coefficients of linear discriminants". That is the direction of Fisher's projection.

Task 4: Add another $n_3 = 150$ observations into your dataset. These group of observations are generated from a bivariate normal distribution with $\boldsymbol{\mu}_3 = (-1, -3)^\top$ and the same covariance matrix as the task 1. Assign the label "3" to those new observations. Next, we understand how does 'lda' function work for a multiple-class problem. We do as follow:

1. Visulaize all the observations in a scatter plot.

2. Apply the 'lda()' function to the full dataset (three groups) and save the output into 'Mod'.

3. Apply the following code

```
x1 <- seq(-6,4,0.1)
x2 <- seq(-6,6,0.1)
d <- expand.grid(x1,x2)
names(d) <- c("V1","V2")
```

   to generate a data frame. This data frame contains all the points of a dense grid of the plane in the scatter plot.

4. Classify each point in 'd' by using the 'predict()' function.

5. Add all points in the data frame 'd' into the scatter plot in step 2. The color of each point is determined by the predicted label in step 4. Now you can find the decision boundary determined by the 'lda' outputs. **Tips**: 'points()' function can be applied to add points into an existing plot.

6. Using the same approach as task 2, draw three decision boundaries (group 1 V.S. group 2, group 2 V.S. group 3, and group 1 V.S. group 3) in the scatter plot. Note: the pooled estimation of sample covariance matrix should be estimated from all the observations (three groups).

What is the relationship between the decision boundary by the 'lda()' outputs (in step 5) and the three lines (in step 6)? Think about it.

Part II: Theoretical problems. Solve the following theoretical exercises: 11.1 and 11.6 in Applied Multivariate Statistical Analysis (AMSA).

Ex 11.1. Consider the two data sets

$$\mathbf{X}_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix} \text{ and } \mathbf{X}_2 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix}$$

for which

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \bar{\mathbf{x}}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

and

$$\mathbf{S}_{pooled} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Assuming the two groups share the same proportion and no cost difference between two type of mistakes, write down the decision boundary of LDA model. (b) Classify the observations $\mathbf{x}_0 = [2, 7]^\top$ as group 1 or 2.

Ex 11.16 Suppose **x** comes from one of two populations:

$$\pi_1 : \text{Normal with mean } \boldsymbol{\mu}_1 \text{ and covariance matrix } \boldsymbol{\Sigma}_1$$

$$\pi_2 : \text{Normal with mean } \boldsymbol{\mu}_2 \text{ and covariance matrix } \boldsymbol{\Sigma}_2$$

Assuming the two groups share the same proportion and no cost difference between two type of mistakes, show that if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then the decision boundary of LDA model is

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$