11. Assignment NMST 539

Frantisek Helebrand

May 2022

1 Introduction

In this assignment we will be analyzing the data from the US, specifically, we will be focusing on the crime data from different US states in 1985. We have observations from 50 different US states with 11 variables.

Let us define a new variable "wd" which split the data into groups according to the region. If the region is Midwest or West we will denote wd as "West". If the region is South or Northeast we will denote wd as "Other". Wd is a binary variable. In this assignment, we will only focus on two additional variables - murders and rape. Now we can go for the discriminant analysis classification.

2 Linear discriminant analysis classification

In this section, we have executed the LDA method for our data. We can see the histogram for the two groups (West and Other.) We can see that the histogram 1 is more spread for group Other also it is centred around -1. On the other hand, the histogram for group West is less spread and centred around 1.

The result (a division of the space) can be seen in the figure 2. The red points are the wrong predictions. The prediction error is 24 % . The yellow and purple point is the mean value for each group.

3 Quadratic discriminant analysis classification

We have also tried another method for computing discriminant analysis. Instead of linear, we assume quadratic discriminant analysis. The result can be seen in the figure 3. The prediction error increased to 26 %. This is caused by the observation in the left-bottom corner. The yellow and purple point is the mean value for each group. If we compare both figures, it can be seen that the areas are quite similar so I would rather choose the linear discriminant analysis (simpler, smaller prediction error.)

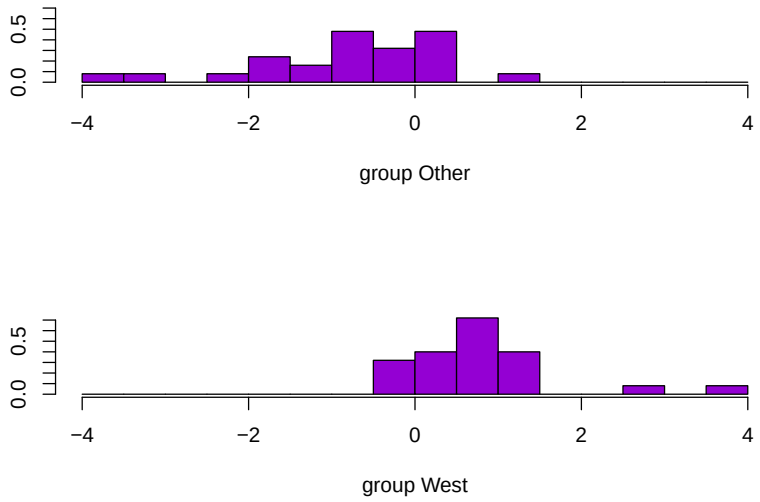

Figure 1: Histogram for group "west" and group "Other".

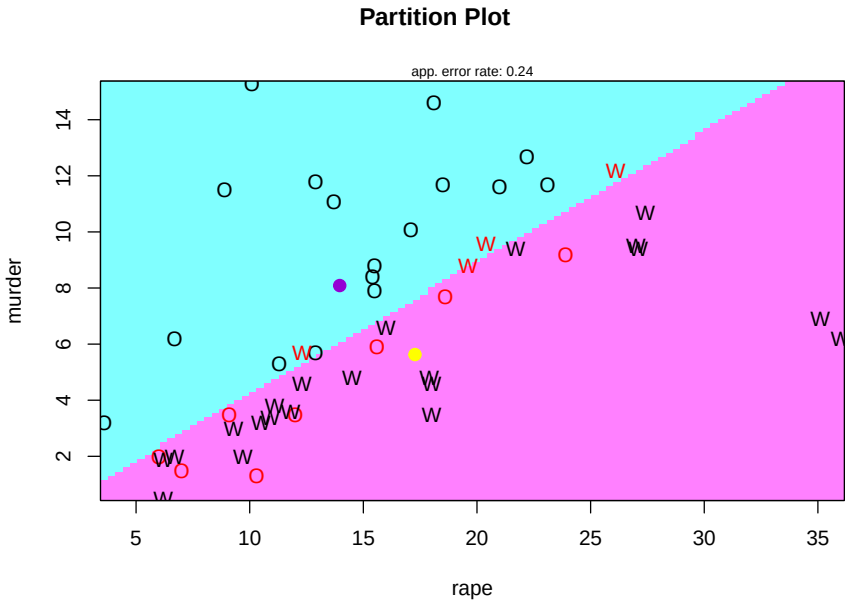

Figure 2: Linear discriminant analysis classification.

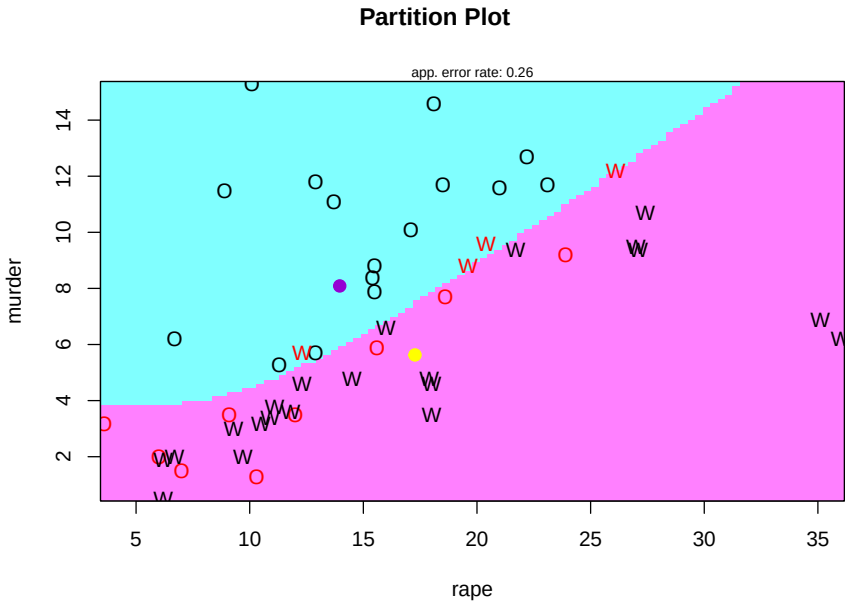

Figure 3: Quadratic discriminant analysis classification.