**Exam of Applied Statistics 15/06/2020**

**Exercise 2**

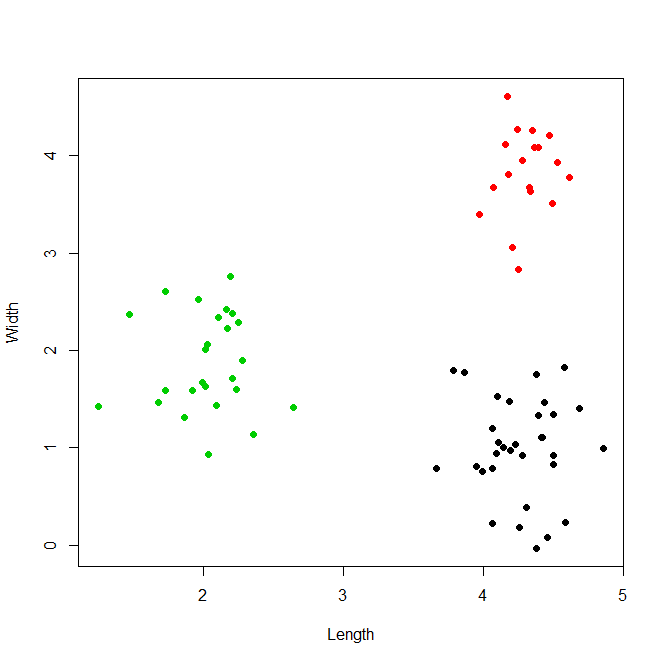
We give a glance at the datas and we can see that maybe there are three clusters of data.

We compute the distance matrix for the datas using Euclidean distance, then using a ward linkage we build the dendrogram here:

Immagine che contiene uccello

Descrizione generata automaticamente

The clustering is not bad, we have a cophenetic coefficient of 0.8898677 and visualizing the datas we can see that it performs well in separing the clusters we have seen before:



Clusters have dimensions 32, 18 e 25. The centroids are (Length ; Width) :

1) 4.261888 ; 1.000128 2) 4.298956 ; 3.827972 3) 2.021256 ; 1.872004

We check that the datas belonging to a certain cluster are normal bivariate by performing a shapiro test for each cluster of datas: we have pvalues high enough (respectively: 0.9336, 0.8428 and 0.3928) to confirm the hypothesis of normality. We also assume the same covariance structures between clusters.

We perform a Manova using the cluster labels: the Wilks statistic suggests us that there’s a sensible difference between the groups, the pvalue is nearly 0 (< 2.2\*10^(-16)). So the membership to a cluster rather than another makes the difference.

We have to build 6 Bonferroni interval for the differences in mean of global confidence 90%, to do so we extract the W matrix from the Manova values and compute the sample means of each cluster. We computer the quantile of the t of confidence 1-0.1/2\*6 and n – g degrees of freedom.

cluster1-cluster2

inf12 mid12 sup12

Length -0.4448921 -0.03706806 0.370756

Width -3.6588441 -2.82784410 -1.996844

cluster1-cluster3

inf13 mid13 sup13

Length 1.871151 2.2406315 2.6101120

Width -1.624745 -0.8718759 -0.1190064

cluster2-cluster3

inf23 mid23 sup23

Length 1.849815 2.277700 2.705585

Width 1.084091 1.955968 2.827845

We can see that the first two clusters (red and black) are not distinguishable by length (we cannot assume the difference of their means to be 0) as we previously see in the picture. For all the other differences of values instead we can see a clear difference from 0, only the difference mean1 – mean3 is a little closer to zero in the Width but still we can say that is different from 0, since its interval does not even contain the zero. These intervals confirms the output from the Manova and also our previous visive analysis. Giving a glance at the coloured plot of the clusters we can see how this intervals are coherent with the distances among clusters that we can see there.