STA 6707: HOMEWORK 3

Due in class: Tuesday April 5

Problem 1:

This data set consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. For further information on this data set see Chapter 10 (pages 176-189) of the book by J. Zupan, and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*. There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria and Sicily), two from Sardinia (inland and coastal) and 3 from northern Italy (Umbria, East and West Liguria).

The variables are described next:

(1) Region (1=South, 2=Sardinia, 3=North)
(2) Area (1=North Apulia, 2=Calabria, 3=South Apulia, 4=Sicily, 5=Inland Sardinia, 6=Coastal Sardinia, 7=Umbria, 8=East Liguria, 9=West Liguria
(3) palmitic acid (
(4) palmitoleic acid
(5) stearic acid
(6) oleic acid
(7) linoleic acid
(8) linolenic acid
(9) arachidic acid
(10) eicosenoic acid

(i) Use the Italian olive oil data set to cluster the observations, without using the variables Area and Region. Comment on the quality of the clustering solution obtained.

(ii) Comment on the agreement/disagreement of your solution to the original assignment of observations to Area and Region.

(iii) Using your clustering solution create a class variable y. Use this new data set to construct two different classification rules. Comment on the quality of the results obtained.