**STA 6707**

**HOMEWORK 3**

**DUE IN CLASS THURSDAY MARCH 21**

**Problem 1:**

Install package pdfCluster and extract the oliveoil data set, using the command data(oliveoil)

1. This data consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. (An analysis of this data is given in (Forina, Armanino, Lanteri & Tiscornia 1983)).

There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria).

The data available are:

Region: South, North or Sardinia

Area: Sub-regions within the larger regions (North and South Apulia, Calabria, Sicily, Inland and Coastal Sardinia, Umbria, East and West Liguria

Palmitic Acid Percentage ×100 in sample

Palmitoleic Acid Percentage ×100 in sample

Stearic Acid Percentage ×100 in sample

Oleic Acid Percentage ×100 in sample

Linoleic Acid Percentage ×100 in sample

Linolenic Acid Percentage ×100 in sample

Arachidic Acid Percentage ×100 in sample

Eicosenoic Acid Percentage ×100 in sample

Build two classification models: one where the class variable corresponds to Region (3-class problem) and one where it corresponds to Area (9-class problem).

Comment on the results (cross-validated error rates, which variables are important if the classifier allows you to do variable selection)

1. Using the model for Area, how well can you predict Region?