STA 6707: HW 2

Due in class Thursday February 25

Problem 1:

Train at least four classifiers for the wine data set from HW 1.

Discuss how you assessed their quality.

Problem 2:

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 (% in sample) (4) palmitoleic acid 2 (5) stearic acid (6) oleic acid (7) linoleic acid (8) linolenic acid (9) arachidic acid (10) eicosenoic acid

Use the Italian olive oil data set to construct classification rules for (i) Area and (ii) Region. Specifically, use at least 3 different techniques from the following list: LDA, QDA, decision trees, random forests and NN. Comment on the results.

Use these classification rules to classify the observations in the file olive-test.dat.

Comment on the (dis)similarity of the rules constructed for Area and Region.**:**