STAT 575 – Assignment 4 – due date is on course outline

For each of questions 2,4,5 – please e-mail me your R programs. The three files should be named `yourname_Qx.R`*, where* $x \in \{2, 4, 5\}$*. I should be able to run them exactly as I receive them, and see the output nicely displayed. For these three questions your written assignment then need contain only the discussions of your solutions.*

1. Establish (21.1) in the class notes.

2. The admissions offices of a business school has prepared data on applicants – their GPA and GMAT scores, and their eventual classification as 'admit', 'do not admit' and 'borderline' – populations 1, 2 and 3 respectively. The data are in T11-6.DAT, and are discussed further as Example 11.11 in the text.

   (a) In each of the three groups, assess the marginal and bivariate normality. Comment.

   (b) Assess the assumption that the three populations have the same covariance. Comment.

   (c) Plot GMAT versus GPA in the training sample, using different symbols for each of the three groups. Plot as well the location (3.21, 497) of a new applicant. How does it appear he/she should be classified?

   (d) Carry out a linear discriminant analysis (using prior probabilities estimated from the sample proportions); assess by presenting both the apparent error rate and the estimated rate using Lachenbruch's holdout procedure. Exhibit the confusion matrices as well. How should the new applicant be classified? Does this change if equal priors are used?

   (e) Repeat (d), using quadratic discrimination.

   (f) Repeat, using logistic discrimination (for which prior probabilities are not an issue). [Use `newdata = rbind(c(3.21,497))` to classify the new observation. In applying the holdout procedure with the multilogistic function, I think you will have to use '`scoring = T`' when observation #59 is being held out – but not otherwise, since Scoring is much slower than Newton-Raphson.]

   (g) Apply Fisher's method with equal priors. Prepare a panel of three plots. The first will be that in (c), the second will be the first sample discriminant, and the third will be both sample discriminants. In each case indicate the position of the new applicant. Classify him/her, using one and both sample discriminants.

... over

3. (a) Establish (23.1) in the class notes – that if a similarity matrix $\mathbf{S}_{n\times n}$ is nonnegative definite (i.e. is a correlation matrix), then there are points $\{\mathbf{q}_i\}_{i=1}^n$ for which (23.1) defines a distance between $\mathbf{q}_i$ and $\mathbf{q}_j$.

   (b) Establish (23.2) in the class notes.

4. Consider the data set on brands of cereals in T11-9.DAT and discussed in Exercise 11.34 of the text. There are measurements on calories, protein, fat, sodium, fibre, carbohydrates, sugar and potassium. (The 'groups' in the final column are just proxies for the manufacturers. The brand names are in the first column.)

   (a) Carry out centroid linkage and complete linkage agglomerative hierarchical clustering. Form 4 clusters in each case. Show the dendograms.

   (b) Carry out K-means clustering, with $K = 4$.

   (c) Carry out model-based clustering, forming 4 clusters and assuming a mixture of normal densities.

   (d) Present the clusters for all 4 methods. Comment on any noticeable similarities and/or differences among the results. Are there cereals that are substantially different from the rest? Is there a relationship between clusters and manufacturers?

5. Investigate the use of multidimensional scaling, attempting to inform your answers in 4-(d).