Statistical Data Analysis
Exercise Session 7

1. Load the data from cafedata.txt. This contains a data matrix with 16 regressors and the response variable "Sales" (these are data from a university cafe). With the first 30 observations:

   (a) Compare the different criteria (R^2, R_p^2, C_p and AIC) to select the best variables from X for explaining y.

   (b) Draw figures of R^2, R_p^2, C_p and AIC as a function of the number of regressors p. Choose the best model based on each criterion.

   (c) Calculate the cross-validated PRESS value for each of these best models. Which model delivers the smallest PRESS value?

   (d) Calculate the mean squared prediction error (MSEP) based on observations 31 to 47 for each of these best models. Which model delivers the smallest prediction error?

2. Consider the same dataset as in exercise 1. With the first 30 observations:

   (a) Perform stepwise regression, starting from the smallest model, with forward search and stepwise search. Also perform stepwise regression, starting from the largest model, with backward search and stepwise search. Compare all models and also compare with the models from the previous exercise.

   (b) Calculate the MSEP based on observations 31 to 47 for each of these best models.

3. Import the weather dataset. The columns are respectively wind-speed, temperature and dew-point temperature measurements recorded in 160 stations around Windsor on "2013-01-16 05:00:00 GMT" (obtained from NOAA).

   (a) Would you scale these variables? Perform the rest of the exercise on the version of the dataset you think is most appropriate.

   (b) Perform a kmeans analysis of this dataset. How many clusters would you retain? Explain why.

   (c) Perform a kmedoids analysis of this dataset. How many clusters would you retain? Explain why.

   (d) Compare the results of the kmeans clustering with the kmedoids clustering. Which one seems best? Explain your choice. Use the clusplot function to visualize the resulting clusters.

   (e) Now perform the agnes algorithm with single, complete and av- erage linkage method. Compare the results with the diana algorithm by making the hierarchical tree. Do the results correspond with the clusters obtained with the partitioning methods?

   (f) Make a three-dimensional plot of the data in which your color the observations according to the clusters. Compare the clustering with the actual labels (which are in the file Lwindsor.txt).

4. Import the skull dataset. This dataset was originally collected by Colonel L.A. Waddell in south-eastern and eastern Tibet, see Morant (1923). The variables are:

   - Length: greatest length of skull,
   - Breadth: greatest horizontal breadth of skull,
   - Height: height of skull,
   - Fheight: upper face height,

- Fbreadth: face breadth, between outermost points of cheek bones.

(a) Answer questions (a):(e) from the first exercise but using the Tibetan Skull dataset.

(b) You can compare the results of the clustering approach you deemed best with the actual labels (in the file Ltibet.txt)

5. Consider the first 9 variables of the pottery dataset (these are measurements of the concentration of various chemical components done on 45 antic pots found in 5 different sites across the U.K.).

   (a) Answer questions (a):(e) from the first exercise but using the British pottery dataset.

   (b) You can compare the results of the clustering approach you deemed best with the actual labels (in the file Lpottery.txt).

Some useful functions:

1. regsubsets() in the leaps package
2. The functions dist, kmeans, pam, agnes, diana, clusplot, silhouette in the cluster package
3. The plot3d function in the rgl package
4. xtabs to compare two sets of labels: xtabs(reallabels+predictedcluster)
5. stepAIC() in the MASS library.