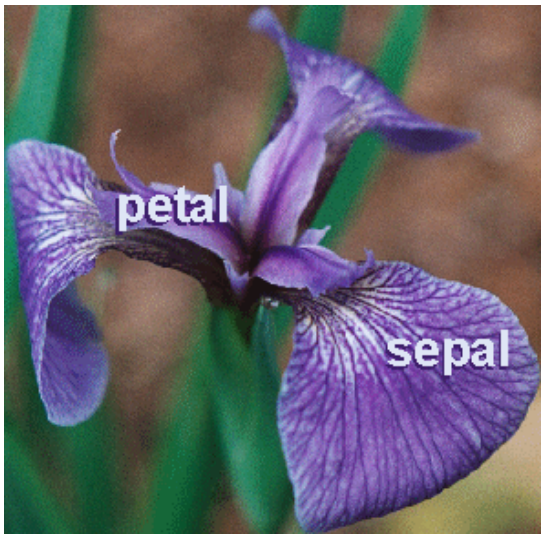
STAT3006 Assignment 2

Due: 17/9/2021 ; Weighting: 30%

Exploratory Data Analysis, Hypothesis Testing and Clustering

This assignment will focus on clustering, but also involves some exploratory data analysis and multivariate hypothesis testing.

At first we will focus on the famous Iris dataset, analysed by Ronald Fisher and many others. The dataset contains 150 observations, each recording 4 measurements of lengths of parts of an Iris flower: the length and width of a sepal and the length and width of a petal. The sepals enclose the flower before it opens, but remain under it after opening. In the case of the Iris, they are much more obvious than the petals. The measurements were made on fully open flowers and are in cm. There are many species of Irises, and 50 observations were collected from three species known at the time: Iris setosa, Iris virginica and Iris versicolour.



More information on this dataset can be found at: http://en.wikipedia.org/wiki/Iris_flower_data_set

The dataset is immediately available in R, just by typing **iris**. You can access the measurements via e.g. iris$Sepal.Length. The species labels are stored under iris$Species.

We assume that each species is equiprobable in the environments of interest and that our data was collected by random sampling from such a population. Neither is strictly true, but we can view the sample as representative and the prevalence of each species is similar in some environments. (See section VI of Fisher, 1936 for some details on how the observations were collected.) This dataset will be used for both clustering and classification (in the next assignment).

Later parts of the assignments include some theoretical work and clustering of both the Iris data and an artificial two-dimensional dataset stored in artificial2021.csv. In answering each question, give some justification and explanation.

Tasks

1. Exploratory data analysis and basic modelling of the Iris Data

(i) Produce a set of bivariate plots, including all possible combinations of two explanatory variables and colour the observations according to their species or plot different species on different plots. Any colour choice is ok provided that they can be distinguished when printed in black and white. It may be useful to use different symbols for each species. The pairs command in R is one option. The plot should be given a number and a caption containing sufficient information to understand it in isolation (i.e. likely more than one sentence). [1 mark]

(ii) Find the (sample) correlation between each type of measurement for each species – report as correlation matrices. Detail any measurements and species for which the absolute value of the sample correlation is 0.7 or greater. Also find two attributes within any of the species for which the sample correlation is not significant at the 0.05 level and give details. [1 mark]

(iii) Try to determine if these classes are multivariate normal or not and explain any method used, including a reference. Comment on possible effects of non-normality in hypothesis testing and clustering. [1 mark]

(iv) Plot sample marginal distributions for each dimension for each class (e.g. kernel density estimates). Fit a multivariate normal distribution to each class using maximum likelihood estimation, noting any caveats. Using the fitted distribution, determine and report the marginal distribution (mathematical form and parameters) for each dimension for each class. [1 mark]

(v) For the virginica class in the Iris data, using the fitted distribution, determine the conditional distribution for the sepal length and width, conditioned upon the (sample) mean values for petal length and width. Produce a contour or perspective plot to illustrate this distribution. Also do the reverse - produce a distribution and contour plot for petal length and width conditioned on (sample) mean values for sepal length and width. [1 mark]

(vi) Think of and explain a way to include a representation of the observations on the plot of a conditional distribution, realising that you can only show two dimensions. Use and justify this method (of adding points to a plot of the conditional distribution), mentioning strengths and weaknesses. [2 marks]

(vii) Produce a residual matrix ($2*n_g$) for each class for petal length and width, conditioning on sepal length and width (n_g = number of observations in the gth class). For each class, plot the residuals as points on a bivariate plot (separate plots will probably be clearest). Comment on whether or not the residuals appear (bivariate) normally distributed within each class. [1 mark]

(viii)

(a) Determine the Mahalanobis distances between each pair of species, and discuss the reasonableness or otherwise of any assumptions. [1 mark]

(b) Which two species seem the most similar - explain why? [1/2 mark]

(c) Which two attributes seem to discriminate best between the three species? [1/2 mark]

The following example R commands can get you started.

```
attach(iris) # attach iris data “to the search path”, i.e. make it available directly for commands
mean(iris[iris$Species=="setosa",1:4]) # will work without attach(iris)
mean(iris[Species=="setosa",1:4]) # needs iris to be attached
tapply(iris$Sepal.Length,iris[5],"mean")
apply(iris3,2:3,"mean") #iris3 is another version of the iris dataset, stored as a 3D array
attributes(iris)
attributes(iris3)
sd(iris[Species=="setosa",1:4])
cor(iris[Species=="virginica",1:4])
cor.test(iris[Species=="setosa",1],iris[Species=="setosa",2])
```

2. Hypothesis testing

(i) Test whether or not there is any difference in the (multivariate) means of the two species Iris versicolor and Iris virginica at a 0.05 significance level. You should use a test such as Hotelling’s T^2 test to do this. Give mathematical details of the test, then give and discuss the results. [1 mark]

You can access a Hotelling T^2 test in R via the manova procedure (see ?manova in R) or a number of other packages.

(ii) The power of the Hotelling T^2 test is weaker for smaller sample sizes. Show the effect of this via a graph of the p-value versus sample size for the comparison above, after choosing random subsets of the Iris versicolor and Iris virginica samples, at least down to a sample size where the null hypothesis is retained. [2 marks]

3. Clustering

(i) Derive the EM algorithm for a multivariate normal mixture model with *H* components, with common covariance matrices. That is - derive the E and M steps for the update equations for the means, covariance matrix and proportion parameters. To do this you will likely need basic Lagrange multiplier optimisation (for the mixing proportions) and some matrix and vector derivative results (see e.g. Seber (2008), Petersen and Pedersen (2012) or Magnus and Neudecker (2019). [6 marks]

Hints: You will need to define the Q function and any notation used. Where possible, use the same notation as present in the lecture notes, but define everything. To derive the M step, you will need to set the derivative of the Q function to zero (including vectors or matrices of zeroes, where necessary) with respect to a parameter type, e.g. a component mean or the common covariance matrix.

We suggest you start with a vector of the mixing proportions, π, leaving the remainder of the unique parameters in a vector θ. We know that the mixing proportions must sum to 1. Lagrange optimisation suggests that you set a Lagrangian function as follows, with λ being a Lagrange multiplier.

$$\Lambda = Q\left(\pi, \theta \middle| \pi^{(t)}, \theta^{(t)}\right) + \lambda\left(\sum_{j=1}^{H} \pi_j - 1\right)$$

You then set its derivative to 0 with respect to e.g. π_k, the *k*th component proportion, and solve for π_k. You then use the sum constraint and solve for λ. Note that you can assume you have current estimates of the τ terms at this point from the recently completed E step. This should lead to an equation for $\pi_k^{(t+1)}$, i.e. the M step for this parameter.

Other M steps should not need Lagrange optimisation, but they will need matrix or vector derivatives.

(ii) It is often claimed that a Gaussian mixture model with spherical covariance matrices ($\Sigma_h = \mathrm{a}_h \mathrm{I}_p$, where $\mathrm{a}_h > 0$) is the same as K means clustering with the same number of components. However, this is not quite true. Explain the ways in which this is not true and what constraints on a mixture model and changes to the EM algorithm would be needed to make this close to true in practice, while retaining the data-generating capabilities of the mixture model. [3 marks]

(iii) Describe with mathematics and pseudocode how you propose to choose the number of clusters for a real dataset with both K means and mixture models. [1 mark]

(iv) Perform exploratory data analysis for the artificial dataset to try to determine how many clusters might be present. Argue for this number, supported by any relevant plots and/or numerical summaries. [1 mark]

(v) Apply K means and mixture model algorithms to both the iris dataset and the artificial dataset (see Blackboard). Comment on the number of clusters chosen and any form of uncertainty about that number – did it agree with what you were able to see from the exploratory data analysis? [2 marks]

(vi) Give parameter estimates for each form of clustering. Include approximate 95% (marginal) confidence intervals for all of the mixture model parameters for the artificial dataset. If using a re-sampling approach, look for evidence of label switching and comment on why it was or wasn't present. [2 marks]

(vii)

(a) Produce a contour plot of the overall fitted mixture distribution on the artificial data. [1 mark]

(b) Produce contour plots of the components of the fitted mixture distribution on the artificial data. Use the same set of weighted density levels for each component. [1 mark]

We prefer that you use the R software environment for this assignment. This is available on all computers in the Maths Department and is also free to install on any of your own computers. Information and downloads are available from http://www.r-project.org/ . R studio (free version) is a recommended integrated development environment for R, available at https://www.rstudio.com/ .

You need to submit two files for this assignment to Blackboard. The first should be a report which answers the questions above, including any graphs, tables, equations and references. This should be saved as a pdf file, which is easy to do from Latex (recommended), Lyx or Word.

The second file should contain any code or scripts that you have written as part of completing the assignment. This could be in a single text file or a set of text files collected in a zip file. There are other options, but the focus should be on the code. Data and output should not be included.

Please name your files something like Yourgivenname_Yourfamilyname_STAT3006_A2.pdf or similar with your student number to make marking easier.

You should not give any R commands in your main report, although you should mention which major libraries you used. Your report should not include any raw output – i.e. just include figures from R (each with a title, axis labels and caption below) and put any relevant numerical output in a table or within the text.

As per https://my.uq.edu.au/information-and-services/manage-my-program/student-integrity-and-conduct/academic-integrity-and-student-conduct, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. You should use consistent notation throughout your assignment and define all of it.

Some references:

Anderson, T.W. *An Introduction to Multivariate Statistical Analysis* 3rd ed., Wiley, 2003.

Duda, R.O., Hart, P.E. and Stork, D.G., *Pattern Classification*, 2nd ed., Wiley, 2001.

Hardle, W.K. and Simar, L., *Applied Multivariate Statistical Analysis*, 4th ed., Springer, 2015.

Magnus, J.R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd ed., Wiley, 2019. https://onlinelibrary-wiley-com.ezproxy.library.uq.edu.au/doi/book/10.1002/9781119541219

Maindonald, J. and Braun, J. *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd ed., Cambridge University Press, 2010.

McLachlan, G.J. and Peel, D. *Finite Mixture Models*, Wiley, 2000.

Morrison, D. F. *Multivariate Statistical Methods*, 4th ed., Duxbury, 2005.

Petersen, K.B. and Pedersen, M.S. *The Matrix Cookbook*, 2012. https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Seber, G.A.F. *A Matrix Handbook for Statisticians*, Wiley, 2008. https://onlinelibrary-wiley-com.ezproxy.library.uq.edu.au/doi/book/10.1002/9780470226797

Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S*, 4th ed., Springer, 2002.

Wickham, H. and Grolemund, G. *R for Data Science*, O'Reilly, 2017.