Midterm Exam, Multivariate Analysis, Spring 2024

2024/4/24

Notes

1. You have three hours to complete this exam. Show all relevant work: partial credit will be given.
2. Gather all your answers into an R Markdown (.Rmd) file. Then, upload this R Markdown file and its compiled html or pdf file to e3 (https://e3.nycu.edu.tw/) under the folder “midterm”.
3. If you have written part of your answers in the answer book, please also upload the scanned file of your written answers.
4. You can use “STAT-Wireless” to gain access to the internet.

Good luck!

Questions

The dataset *oliveoil* contains 572 rows, each corresponding to a different specimen of olive oil, and 10 columns. The first and the second column correspond to the macro-area (Centre-North, South, Sardinia) and the region of origin of the olive oils, respectively. Columns 3-10 represent the following 8 chemical measurements on the acid components for the oil specimens: palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic. The data set *oliveoil* can be downloaded from:

https://ghuang.stat.nycu.edu.tw/course/multivariate24/files/exam/oliveoil.csv (https://ghuang.stat.nycu.edu.tw/course/multivariate24/files/exam/oliveoil.csv)

The data set can also be downloaded from e3 (https://e3.nycu.edu.tw/) under “midterm”.

In this exam, you will be asked to perform various multivariate analyses on this data set using the R software.

Question 1

To examine the differences of acid chemical measurements across three macro-areas, one can do the multivariate mean inferences.

a. (15 points) Use the one-way MANOVA to examine the overall acid chemical measurement differences among different macro-areas. Write out its MANOVA table for comparing population mean vectors. What is the null hypothesis of this test? What do the test results tell you?

b. (15 points) Also, perform the one-way ANOVA on each each acid measurement (8 variables in total) for its differences over macro-areas. Since we need to perform the test for multiple measurements simultaneously in the ANOVA analysis, what is the cut-off for the p-value, below which can be considered to be significant? Which acid measurement(s) are significantly different over macro-areas?

Question 2

(30 points) Now, you will be asked to perform the principal component analysis (PCA), the orthogonal factor analysis (FA) with a proper factor rotation, and the multidimensional scaling (MDS).

Plot all observations in the *oliveoil* data set by their 1st and 2nd principal components from PCA. Also, plot them in terms of their 2 factor scores of FA. The two-dimensional representation produced by MDS is requested. In these figures, you should use different colors for observations from different macro-areas. What do these figures tell you about the closeness of three macro-areas on acid measurements?

Question 3

Do (1) the agglomerative hierarchical clustering with average linkage, (2) the k-means clustering, and (3) the model-based clustering that adopts the Gaussian mixture model with covariance matrices $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_3 = \mathbf{\Sigma}$.

a. (10 points) Which approach has the best performance in clustering specimens from the same macro-area together?

b. (10 points) From the results of k-means clustering, what are the (total) within-cluster sum of squares and the between-cluster sum of squares?

c. (10 points) For model-based clustering, write out the Gaussian mixture model used for clustering. Please specify the estimated values of the parameters in the model.

d. (10 points) To assess cluster fit, you are asked to create the silhouette plot for each of the clustering approaches. Which approach has the best cluster fit based on the average silhouette width?