IE 582 Statistical Learning for Data Mining
Homework 3
Due January 20, 2024

Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

a) A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
   - https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
   - https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps917519326e41

   Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

b) A Markdown html document. This can be created using RMarkdown for R and Python. Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself.

Please mention if you utilize Large Language Models (i.e. chatGPT, BART and etc). You are expected to provide your prompt and resulting response of these models at the end of your reports as Appendix if you have utilized any of these tools.

As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Clustering and Estimation of Mixture of Multivariate Normal Distributions

Random forest [1] is one of the tree-based ensemble learning approach which is claimed to model a mixture of multivariate Gaussian distribution. In other words, they can be used for infer density information. Simulate the following experiment to see if this discussion holds:

1- Generate 500 data points with 8 variables with 4 different parameter settings (i.e. covariance and mean). You are expected to have 500x4=2000 data points with 8 variables in total.
2- Train an unsupervised random forest to find the random forest similarity (or proximity [2]) between 2000 observations.

3- Transform similarity to dissimilarity (with an appropriate transformation) and use two alternative clustering approaches to obtain four clusters: partitioning around medoids clustering algorithm and hierarchical clustering with Ward's method. Moreover, apply k-means clustering to raw data to obtain four clusters so that you can compare to a baseline approach.
4- Compute sample mean vector and sample covariance matrix for each cluster from three clustering strategies and compare with the parameters you used to generate the synthetic data.
5- Compare the performance of the proposed strategies when we introduce additional noise variables from Bernoulli distribution up to 8 variables (i.e. as many as original number of variables).

Briefly discuss why random forests work (or do not work) for this type of problems. How would the performance change with the increase in the dimensionality (i.e. number of variables)?

References:
[1] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[2] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox