Project 4: Clustering

Objectives

To use various clustering algorithms to determine the best number of clusters in a data set that will be created using your student id (SI) number. For this analysis you can use any statistical package, such as MatLab, R, SAS, or use Python with all the available statistical functions.

Data Set

You will use the data set provided to you that is completely randomized based on your student id. The data set consists of about 600 3-tuplet observations.

Tasks

1. *Hierarchical clustering*
   Apply the hierarchical algorithm to the dataset.
   1.1 Plot the dendrogram and the distance graph (if it is given by your package)
   1.2 Determine the number of clusters.
   1.3 Color the data according to their cluster, and do a 3D scatter diagram. Rotate the diagram to identify the clusters.
2. *k-means clustering*
   2.1 Apply the algorithm for several values of k starting with k=2.
   2.2 Use the elbow method to determine the best value of k.
   2.3 For the best k value, color the data according to their cluster, and do a 3D scatter diagram. Rotate the diagram to identify visually the clusters.
3. *DBSCAN clustering,*
   Apply the DBSCAN algorithm to the dataset to determine the number of clusters.
   3.1 For Minpts=3, use the elbow method to determine the best values of ε. Run the DBSCAN algorithm for the best value of ε and Minpts=3. Color the data according to their cluster, and do a 3D scatter diagram. Rotate the diagram to identify visually the clusters.
   3.2 Repeat the above step for Minpts=4,5,6. (use more values if necessary).
   3.3 In your report provide only your best clustering and its 3D scatter diagram. Provide the remaining results of your investigation in a separate file.
4. Compare and discuss the results from all three methods. Identify the best clustering of the dataset.

What to submit

For each task, submit your results and conclusions along with the code that you wrote to obtain the results. It is very important that you provide enough results to support your conclusions. Conclusions without insufficient results will make you lose points. Also, it is important that you develop your own code. Sharing code is not allowed and constitutes cheating, in which case both students (the one that aids and the one that receives) will get a zero for the project and will be reported to the student conduct office.

Remember that you will be graded mostly on your ability to interpret the results

Grading
The TA will first verify that your code works and produces the results you submitted. The break down of the grades will be as follows:

Task 1: 30 points
Task 2: 30 points
Task 3: 30 points
Task 4: 10 points

Extra Credit

This is an optional task. It will be graded from 0 to 100, and then transformed into the range [0,3]. The transformed grade will be added to your final numeric grade. It may help you go up one +/- grade category.

Task: Use the Gaussian decomposition method to cluster the same data set you used above.

1. Plot the maximum likelihood against *k* (number of clusters) and select best *k.*
2. Obtain a projection of the fitted mixture of normal distributions on each plane, i.e. *x*-*y*, *x*-*z*, and *y*-*z*, something similar to fig. 8.12. If possible, color the projection to identify each normal distribution, or simply identify the clusters by highlighting the pdf image by hand.
3. Compare your results to those obtained above in project 4 and discuss your findings.

Grading
Task 1: 60 points
Task 2: 20 points
Task 3: 10 points