MALMÖ UNIVERSITY

Technology and society

Written Exam
Exploratory Data Analysis
MA661E, Spring 2025

Solutions should be in reasonably details, and be written in Word/Latex

The codes in Python / Matlab should be attached as appendix

Dead line for hand-in is 2025-03-29, 23:59 via Canvas.

1. Due to large sizes of data sets in practice one need to do dimension reduction thereafter study the pattern by various clusterings techniques. Dimension reduction means lost of certain information whence less accuracy of clustering results in general. In this question, we consider the Iris data set as four dimensional data set with the ground truce label: setosa, versicolor, virginica and use the k-means for clustering.

a) First determine three clusters of Iris data set by k-means, and even compute the percentage of correct classified observations. (4p)

b) Now apply first the factor analysis method to reduce the Iris data set to two dimensional, then determine three clusters of the reduce data set by k-means. What is then the percentage of correct classified observations ? Even visualize the reduced data set with original label respectively by the clustering. (8p)

2. Consider the minimum spanning trees method (MST) and lung cancer data set `LungA`.

a) First apply MST to find two clusters, then find the percentage of correct clustered genes using the original label of cancer cell: `Normal, Cancer (Small-cell lung carcinomas, Nonsmall cell lung carcinomas)` as ground truth. (4p)

b) Apply MST once again to find three clusters, then compare the correct clustered genes using the original label: `Nonsmall cell lung carcinomas, Normal, Small-cell lung carcinomas` as ground truth. (3p)

c) Finally, test to find five clusters comparing the correct clustered genes using the label: `AD, COID, SQ, NL, SCLC` as ground truth. Even comment on these clustering results. (4p)

3. Model based clustering in chapter 6 (section 6.5) is based on geometrical properties of clusters, e.g. balls and ellipsoids in 3-dim. In this question, you are asked to randomly generate data sets to check the capability of MBC: two balls of different sizes, two ellipsoids of different size with symmetry axis parallel with coordinates axis, three ellipsoids also different sizes with arbitrary symmetry axis.

a) Randomly generate those sub data sets so that they are disjoint (far away from each other), then apply model based clustering to fin clusters even compare the estimated clusters with original labels. (7p)

b) Randomly generate those sub data sets so that they are overlap partly, then apply model based clustering to find clusters even compare the associated clusters with original labels. (5p)

4. Find the datasets about inflation, unemployment in member countries of European union during the last ten years (2015 - 2024). Your dataset should contains at least 20 countries. Denote the inflation dataset by $EUin$ and the unemployment dataset by $EUun$ and $EU = [EUin, EUun]$.

Hint You may check the database at the websites: https://european-union.europa.eu or https://ec.europa.eu/eurostat/data/database or

a) Study the unemployment pattern by looking at clusters of $EUun$ using one proper hierarchical methods and evaluate the result by Silhouette plot. (5p)

b) Study the joint pattern of unemployment and inflation by looking at clusters of EU using k-means method and evaluate the result by using Dunn index. (6p)

5. Consider the dataset $X^T = \begin{pmatrix} 3 & 1 & 1 & 4 & 1.5 & 0.12 & 0 & 0.03 & 0.1 \\ 1 & 4 & 1 & 1 & 3 & 2 & 6 & 0.5 & 4 \\ 0.1 & 0.02 & 0 & 0.1 & 0 & 1.9 & 3.5 & 1 & 3 \end{pmatrix}$

a) Visualize the dataset X via `scatter3`. (2p)

b) It's obvious that X contains two clusters: one is on xy-plane and another one is on the yz-plane. Please check if hierarchical Ward's method, spectral method are able to recover these two clusters. (8p)

6. Consider again the dataset X given in the previous question.

a) Determine the non-negative factorization matrices W, H of X, i.e., $X = WH$, by *Multiplicative update algorithm*(MULT). (3p)

b) In this question, we study the convergence of MULT. Let the maximum number of iterations be K and the error function f be $||X - WH||^2$, then f is a function of K. Modify the Matlab EDA toolbox routine `nnmf` or any your Python library such that MULT terminates whenever the number of iterations has reach the maximum number of iterations K.

i) Plot the error term f as an function of K. (5p)

ii) For what integers K is the error f less than the tolerance $\varepsilon = 10^{-4}$? (2p)

Hint The Frobenius norm $||A||$ for matrix $A = (a_{ij})$ is $||A||^2 = \sum_{i,j} a_{ij}^2$.

7. In the study of dimension reduction by Principle Component Analysis (PCA), one of PCA is based on covariance matrix, and another one is based on correlation matrix. The covariance matrix based PCA has a property: all the PC:scores are uncorrelated, i.e., the correlation coefficient between any pair of PC:scores is zero. Please give a short proof of this property. (6p)

8. Consider the Sierpinski triangle, which is a fractal set and is of Hausdorff dimension $\frac{\log 3}{\log 2}$, roughly 1.585.

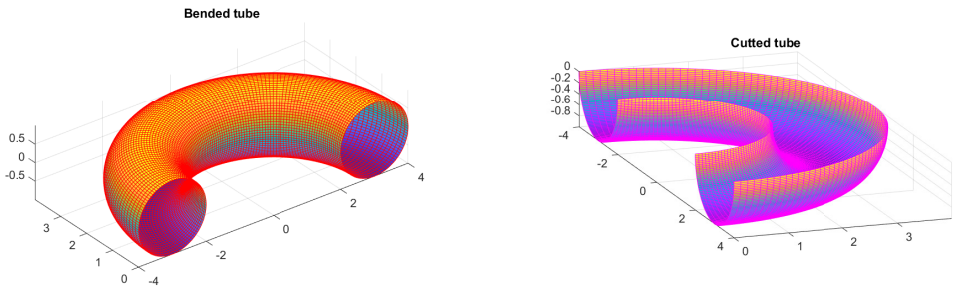
a) Please make samplings of points from the above Sierpinski triangle of sizes: a) 10000 points, b) 100000 points. (2p)

b) Estimate the intrinsic dimensionality of the above Sierpinski triangle by the nearest neighbor method `idpettis` (4p)

c) The number of neighbors K in the above EDA toolbox `idpettis` is by default 5, study the sensitivity of the above estimations with respect to K. Plot these estimates as function of K, where $5 \leq K \leq 15$ and even comment on the behaviors of these curves. (4p)

Hint: You may check the file: sierpinski.m, which is available in Canvas.

9. Consider the bended tube and also the cutted tube (see the figure below). These two surfaces are topologically very different.



a) Generate 20000 and 10000 sample points from the bended respectively the cutted tubes, and denote them by X, and Z. (2p)
Hint: You may check the file: tube.m, which is available in Canvas.

b) Apply the nonlinear dimension reduction method LLE to these datasets X, Z. How does this method work for the bended tube respectively the cutted tube? (4p)

c) Apply even the reduction method HLLE to these datasets X, Z. Does HLLE work better/worse than LLE? (4p)

10. Consider the SMACOF method (see the description in the separated page).

a) First implement the SMACOF algorithm in Matlab or Python. (4p)

b) Test your code for Leukemia dataset as in example 3.2 in the textbook for the choices of parameter $p = 1.5, \quad 2, \quad 7$. Are there some essential differences in the results for different values of p in Minkowski distance ? (6p)

Good luck !

Grade

$0 - 49$p: F	$50 - 59$p: E	$60 - 69$p: D
$70 - 79$p: C	$80 - 89$p: B	$90 - 100$p: A