Data Analytics (774/874)

Post-block Assignment 1

Department of Industrial Engineering

Stellenbosch University

8 October 2021

Responsible Lecturer: Eldon Burger (eldonburger@sun.ac.za)

Deadline: 22 October 2021 @ 23:59

Total: 45 marks

Instructions

1. The assignment tests your understanding of the concepts covered in topics 1, 2, 3.
2. Your assignment should be submitted as **one pdf** document:
    a. Name the document as ???????PBA1.pdf, where you replace the question marks with your student number.
    b. In your final document, provide a heading for every question based on the notation of the document followed by your answer. For example, **Question 1** followed by your answer.
    c. When a question requires you to motivate your answer, provide sufficient reasoning as no marks will be awarded for answers without a correct explanation.
3. **Late submissions cannot be accepted** and no extensions to the deadline will be provided.

Introduction

Refer to the data set *money.csv* available on the online page of the course. The data set *money.csv* contains the measurements of 200, of the 1000 Swiss Franc banknotes. For each banknote the following six measurements were recorded: the length of the note (x_1), the left width of the note (x_2), the right width of the note (x_3), the width of the margin at the bottom of the note (x_4), the width of the margin at the top of the note (x_5) and the diagonal length of the inner frame of the note (x_6). The various measurements are illustrated in Figure 1.

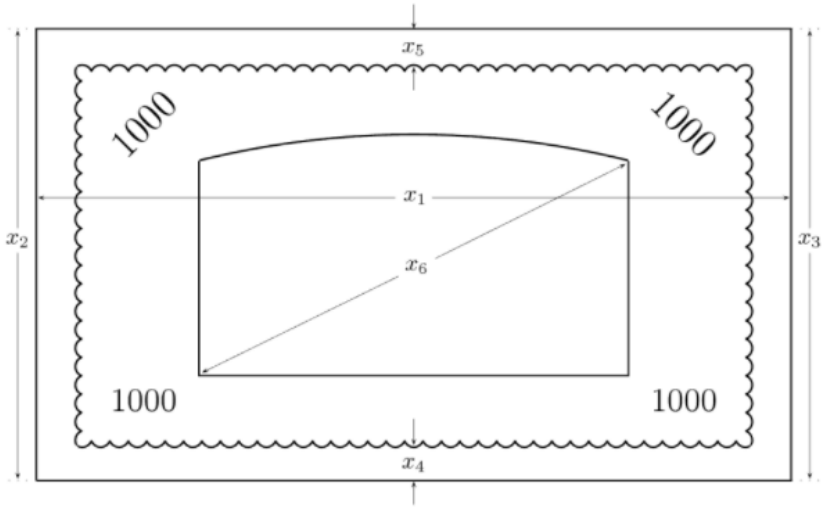

Figure 1: Dimensions of a note

Cluster Tendency

Question 1: Create a scatter plot matrix of the data set; the features used for each individual scatter plot should be indicated on the scatter plot.

[2]

Question 2: Is there any clear cluster visible in one or more of the scatter plots created in question 1? Motivate your answer by referring to one of the specific scatter plots and indicating how many clusters are visible.

[2]

Question 3: Determine whether clustering could produce meaningful results by performing a visual assessment of cluster tendency[1].

Provide:

(i) the ordered dissimilarity images,

(ii) an assessment of the dissimilarity image by explaining whether the ordered dissimilarity image indicates any potential clusters.

[4]

[1] Refer to slide 74 of Topic 2

DBSCAN (Density-based spatial clustering of applications with noise)

Assume that the clustering algorithm DBSCAN has been selected. To apply DBSCAN on a data set the parameters *MPts* and *Eps* must be specified.

Question 4: Given that ***MPts* = 5**, create an ordered k-distance graph[2] to identify suitable values for *Eps*.

[4]

Question 5: From the k-distance graph it is not always clear which specific *Eps* value to select. Instead, the graph can be used to select an appropriate search range for *Eps*. Using the graph provided in question 4, identify a suitable upper (*Eps_max)* and lower (*Eps_min)* search limit for *Eps*. Motivate your selection of *Eps_max* and *Eps_min*.

[2]

DBSCAN parameter search

To determine the *best* values for *Eps*, DBSCAN will be applied to the data set for various *Eps* values and an ***MPts* = 4**. Apply DBSCAN to the data set for eight *Eps* values [0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3]

Question 6: Create a stacked bar chart that shows the number of instances assigned to a different label (coloured by cluster label) for each of the eight experiments. Instances classified as noise points should be excluded from the bar chart.

[4]

Question 7: For each of the eight DBSCAN models, create a scatter plot of the features Diagonal vs Right. Make use of colours to indicate to which cluster each instance is assigned, for each of the models. Use the same colours as used in question 6.

[4]

Question 8: Compute the Davies Bouldin score[3] for each of the models with and without including the instances classified as noise in a separate cluster. Use your answers to populate the table provided below.

Eps	Davies Bouldin (without noise cluster)	Davies Bouldin (with noise cluster)
0.6		
0.7		
0.8		
0.9		
1.0		
1.1		
1.2		
1.3		

[4]

[2] Refer to slide 102 of Topic 2
[3] Refer to slide 79 of Topic 2

Question 9: With regards to the results obtained in question 8, answer the following questions:

- Based only on the Davies Bouldin score, which *Eps* values should be selected when the calculation is performed with the noise instances grouped in one cluster?
- Based only on the Davies Bouldin score, which *Eps* values should be selected when the calculation is performed without the noise instances?
- Is it better to use the Davies Bouldin score with or without considering noise? Motivate your answer

[4]

Question 10: Using the results from questions 7,8,9 select an appropriate value for *Eps*. Create a scatter plot matrix to illustrate the clusters for all pairs of descriptive features, indicate which *Eps* was selected, and motivate your selection of *Eps.*

[10]

Cluster Stability

Question 11: To estimate the stability of the solution provided in question 10, create three subsamples of size 200 with sampling with replacement. Apply DBSCAN to each subsample with the same parameters as used in question 10. Create a scatter plot of the features Diagonal vs Right for each subsample and make use of colours to indicate to which cluster each instance is assigned. Does your solution produce the same clustering results for each subsample?

[5]