ELE 39091 Cluster Analysis for Business
Course paper - 40%

Instructions

Read carefully: Create a single pdf file with your answers, which are limited to 10 content pages, and your Python code at the end as an appendix. Make sure the Python code is readable and split with headers for each question. Your report must contain: i) a front page with all group members' ID as in WISEflow, ii) questions with numbers as header, and iii) page numbers. The front page and appendix are not considered content pages. The answers should be short and concrete and should not include the output of Python functions, e.g. `print(df.head())`, `print(model.summary())`, `print(x)`, or direct screenshots of notebook blocks such as Google Colab. Therefore, you should create your own summary tables in text editors/spreadsheets or generate plots in picture format and embed them into your pdf report. Note that the questions in this term paper are numbered with letters, e.g. a), b), etc., and steps that you should follow are listed with roman numerals, e.g. i), ii), etc. You only need to include the answers to the questions in your report.

Honor Code: By answering this project, I confirm that I will not give or receive, from any person or any AI software, any help in this project, as this is considered cheating. Each group member accepts responsibility for her or his role in ensuring the integrity of the work submitted by the group. **Any suspected cheating or use of AI tools will be reported to the exam administration immediately and students will be called for an oral consultation as an additional verification before receiving a final grade.**

1 Clustering and Representation Learning (50 pts.)

After you completed your bachelor degree at BI Norwegian Business School, Pulpo Data hired you as a junior data scientist. The Chief Analytics Officer, Ms. Marti G., tells you that you will work in the Clustering Analysis team. It is therefore recommended that you revisit the concepts and techniques acquired during your undergraduate studies, utilizing a toy data set for this purpose.

You start by

i) Downloading the data set from

```
path = 'https://raw.githubusercontent.com/milaan9/Clustering-Datasets/refs/heads/master/02.%20Synthetic/banana.csv'
```

and load it onto a `pandas` DataFrame. Then,

a) Visualize the data in a 2-dimensional scatter plot (5 pts).

b) Based on your intuition, explain what the number of clusters in the data should be (5 pts).

Use the Euclidean distance and the ward and average linkage methods and

c) Plot the dendrogram for both linkage methods. Discuss the number of clusters that the dendrograms suggest and contrast the result with your answer in b) (5 pts).

As it is some time since you graduated, you do not remember all clustering techniques that you learned. However, you remember that the popular K-means is a powerful technique, which must be why it is popular after all. So, you

d) Select the K parameter based on your answer in exercise b), run the K-means algorithm, and visualize a 2-dimensional scatter plot where the K-means cluster labels are used as colors in the scatters. Comment on the results (5 pts).

Suddenly, you remember the concept of representation learning and the autoencoder model and want to try it out. Therefore, you use the autoencoder model in the Github repository, but with some modifications. You

ii) Replace the loss function in the original code with

```
self.loss = tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(
    y, x_hat))
```

where `y` is the one-hot encoding version of the `class` variable in the toy data set that you downloaded in step i). You can obtain the one-hot encodings using this code

```
labels = df['class'].copy().to_numpy()
y = tf.keras.utils.to_categorical(labels-1,num_classes=2)
```

iii) Define the `inputs` argument in the `call` function as the tuple `(x,y)`. Hence, the first line below the `call` function is

```
x,y = inputs
```

and, therefore, the argument for `self.encoder` is now `x`. Remember to update the `Dataset` loader accordingly, i.e.

```
tr_data  = tf.data.Dataset.from_tensor_slices((x,y)).shuffle(x.
    shape[0]).batch(batch_size)
```

iv) Add 3 hidden layers in both encoder and decoder with 10 units in each hidden layer and let the dimension of `z` be 2.

v) Use the `'tanh'` activation function in all hidden layers and `'linear'` activations in both output layers.

With these modifications, you train the autoencoder for 1000 epochs using the Adam optimizer with a learning rate of 0.01 and a batch size of 256. Then,

e) Plot the loss across number of epochs and generate the latent representation for all data using the function `rep_learning` and visualize it in a 2-dimensional scatter plot (15 pts).

f) Use K-means again, but this time you will cluster the latent representations that you obtained in exercise f). Visualize the latent representation and use the K-means cluster labels as colors in the scatters. Comment on the results (5 pts).

g) Can you think of a clustering algorithm that is better suited for this data set? Explain the mechanism behind such an algorithm (10 pts).

Note: Training an autoencoder gives slightly different results each time, as it is an iterative procedure. Therefore, it is recommended to train the autoencoder a few times, generate the latent space, and save the latent representation of a successful run (e.g. `np.save('lentent_z.npy',z)`), which you can further use in exercises f) and g).

2 Retail Store Business Case (50 pts.)

Your first project in Pulpo Data is to help a Norwegian retail store understand its customers, their needs, and behaviors. The retail store is currently developing its new strategy and, therefore, is wondering whether it should adapt its product portfolio to better meet its customers. To do this, Marti G. sends you a csv file with actual data that the retail store provided. You find the file in the folder Midterm in Itslearning. The description of the variables are in the Appendix.

You are eager to start working on this project and

i) Load the file into a pandas DataFrame and you called it df

ii) Explore the data. Use any method you think is appropriate. For example, you can use the methods

```
df.head()
df.describe(include='object').round(2).T
df.describe(include='number').round(2).T
```

a) What can you say about the data? Recall that it will be used to identify clusters, or groups of customers, in the retail store (5 pts).

After you explore the data, Ms. Marti G. calls you to a meeting and asks you what your initial thoughts are. Then she suggests that you should

iii) Convert the categorical variable to one-hot encoding using

```
df = pd.get_dummies(df, columns=categorical_variables_as_list,
    dtype=int)
```

where categorical_variables_as_list is a list with the names of the categorical variables.

iv) And *rotate* all 15 variables using PCA. That is, no dimensionality reduction.

Explain

b) What is the reason for using PCA and not reducing the number of dimensions in the input data (5 pts).

In the remainder of the exercise, you use the PCA transformation from iv) and not the input data.

You continue with the difficult step of selecting the number of clusters. That is, the K parameter. You decide to cross-validate K using both hierarchical clustering with ward linkage method and using K-means.

c) For $K = 1, 2, \cdots, 10$, plot the Silhouette and Davies-Bouldin scores for both hierarchical clustering and K-means, i.e. you are supposed to show 4 plots (5 pts).

d) What are the number of clusters suggested by your cross-validation? Does the Silhouette and Davies-Bouldin scores agree with each other? (5 pts)

You were satisfied with the results you obtained with the autoencoder in the toy data set. Therefore, you use the autoencoder in the Github repository with the following modifications:

v) Replace the loss function in the original code with

```
self.loss = tf.reduce_mean(tf.keras.losses.MSE(inputs,x_hat))
```

vi) Add 3 hidden layers in both encoder and decoder with 35 units in each layer and let the dimension of z be 10.

vii) Use the 'tanh' activation function in all hidden layers and 'linear' activations in both output layers.

With these modifications, train the autoencoder for 1000 epochs using the Adam optimizer with a learning rate of 0.01 and a batch size of 256. Then,

e) Generate the latent representation for all data using the function rep_learning and visualize it in a 2-dimensional scatter plot using the t-SNE dimensionality reduction method (10 pts).

f) How many clusters can you identify in the scatter plot? Is it similar to the number of clusters that you identified in exercise d)? Remember the note at the end of the first exercise in page 2 (5 pts).

Ms. Marti G. asks you to write an executive summary of your main findings to be submitted to the retail company. To this end,

viii) Select one of the clustering algorithms in exercise c) together with the optimal K-value found by cross-validation and assign each observation to its corresponding cluster.

Then

g) Write the executive summary for the retail company analyzing each of the clusters. What are the features that distinguish each cluster? Use any method that you consider appropriate, e.g. average values, standard deviations, histograms etc. (15 pts).

Appendix

Variable names of the data `retail_store_data.csv` together with brief explanations.

retail_store_data.csv

```
1. Education:               Customer's education level
2. Marital_Status:          Customer's marital status
3. Income:                  Customer's yearly household income
4. Kids:                    Number of children and teenagers in customer's household
5. Days_is_client:          Number of days since customer's enrollment with the company
6. Recency:                 Number of days since customer's last purchase
7. Expenses:                Total amount spent by the customer
8. CustomerAge:             Customer's age
9. TotalNumPurchases:       Total number of purchases made by a customer
10. TotalAcceptedCmp:       Total number of accepted offers by a customer
11. Complain:               1 if the customer complained in the last 2 years, 0 otherwise
12. Response:               1 if customer accepted an offer in the last campaign, 0 otherwise
```