

Unsupervised Learning Lab

Aim

The aim is not the code in itself but to show an understanding of the underlying concepts and how they are applied to solve a practical problem. This lab is divided into three parts.

To pass the lab, write down your answers to the questions in Part 1, 2 and 3. During the lab demonstration, explain your answers to the lab assistant. The lab assistant will ask you some general questions about the parts of the lab and how you solved them.

You do the lab alone or in a group of two or three persons (not alone).

Part 1

Read the tutorial **K-Means Clustering** on the course website.

Execute the Python scripts of the document and modify the parameters of the algorithm to better understand what happens.

Answer the following questions.

1. What are the relevant features of the Titanic dataset. Why are they relevant?
2. Can you find a parameter configuration to get a validation score greater than 62% ?
3. What are the advantages/disadvantages of K-Means clustering?
4. How can you address the weaknesses?

Part 2

Read first the tutorial **Hierarchical Clustering** and go through the Python code of the tutorial to understand how this type of clustering works.

In this part of the lab, we will perform hierarchical clustering on real-world data and see how it can be used to solve an actual problem. The problem that we are going to solve here is to segment customers into different groups based on their shopping trends.

Download the file `shopping_data.csv` from the course website and place it in your working folder. Open a command (cmd) window and run Python.

To import the shopping dataset execute:

```
customer_data = pd.read_csv('shopping_data.csv')
```

You can explore the dataset a bit. For example, to check the number of records and attributes, execute:

```
customer_data.shape
```

This will return (200, 5) which means that the dataset contains 200 records and 5 attributes. To see the dataset structure, execute the `head()` function of the data frame:

```
customer_data.head()
```

Our dataset has five columns: CustomerID, Genre, Age, Annual Income, and Spending Score.

To view the results in two-dimensional feature space, we will retain only two of these five columns. We can remove CustomerID column, Genre, and Age column. You can retain the Annual Income and Spending Score (1-100) columns. The Spending Score column signifies how often a person spends money in a mall on a scale of 1 to 100 with 100 being the highest spender.

Execute the following script to filter the first three columns from our dataset:

```
data = customer_data.iloc[:, 3:5].values
```

Next, we need to know the number of clusters that we want our data to be split to. Use the *scipy* library to create the dendrograms for the shopping dataset as explained in the Hierarchical Clustering tutorial.

- 1 How many clusters do you have? Explain your answer.

When you know the number of clusters for the shopping dataset, you can group the data points with respect to these clusters. To do so use the `AgglomerativeClustering` class of the `sklearn.cluster` library.

- 2 Plot the clusters to see how actually the data has been clustered.
- 3 What can you conclude by looking at the plot?

Part 3

Read first the tutorial **A One-Stop Shop for Principal Component Analysis**. Then, read the tutorial **Understanding PCA (Principal Component Analysis)** and answer the following questions:

1. Can you choose `n_components=2`? Can you think of some method to test this?
(question at page 6)
2. Create the scatter plot of the third principal component (that is, you combine the third principal component with the first and then the second principal component). What can you see with the plot? What is the difference?
(question at page 7)
3. Can you tell which feature contribute more towards the 1st PC?
(question at page 8)