Lab 3 - Clustering and Manifold Learning (2%)

In this exercise, you will be introduced to standard techniques for unsupervised ML used also for EDA. These include the *k-means* clustering algorithm and *manifold-learning* techniques, such as multi-dimensional scaling (MSD) and t-SNE. This exercise aims to practice these techniques for discovering and visualising hidden patterns in data.

At the end of this exercise you should be able to:

- Apply a clustering algorithm for partitioning a data set into clusters of similar examples.
- Use a manifold-learning technique to visualise a clustered data set with 3 or more numerical attributes as a 2D scatter plot.

Task 1

- Download the following dataset and example notebook:
  - iris_nolabels.csv
  - Lab 3 - Clustering and Manifold Learning.ipynb

Task 2

- Follow the examples in the notebook **Lab 3 - Clustering and Manifold Learning.ipynb** to cluster either the **loans_train** dataset or the **bank** dataset after it has been prepared for ML (i.e. the result of Lab 2). Use the k-means clustering algorithm and visualise the clustering. Pick a value of k suggested by MDS and t-SNE visualisations of the dataset. Experiment with various values of the parameters of k-Means. Aim at producing a meaningful clustering that allows you to describe the clusters with words. Write a short conclusion about the characteristics of the clusters.

Task 3

- Plot the sum of squared distances from the data points to the k-Means clusters' centres for various values of k. Use the Elbow method to pick the best value of k. If it suggests a different value of k (from the one you used in Task 2), compute a new k-Means clustering.

Task 4

- Implement and plot the results of the Silhouette method for determining the best value of k for k-Means. Compare the results of the Elbow method to the results of the Silhouette method.

Task 5

- Apply another clustering algorithm (from the ones available in scikit-learn) to the same dataset and an appropriate manifold-learning technique to visualise the clustering.

Submission

Save your Jupyter notebook and name it *lab3.ipynb*.

Submit both your notebook and the prepared dataset you have used in the Course Tools > Assignment section by 08-Mar-2024 23:59. Late submissions will not be accepted.