ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Machine Learning
Summer 2023

Prof. Dr.-Ing. Klaus Berberich
Phone: +49 681 58 67-243
klaus.berberich@htwsaar.de

Programming Assignment 3

The programming assignment will be discussed on **July 10th, 2023**. To obtain bonus points, you have to submit your solution as a Zip-file via Moodle by **July 7th, 2023 23:59**. Teams of up to **three** students are allowed. Please indicate your names and your student numbers in your submission.

3.1 Clustering News Articles (5.0)

For this programming assignment, we are going to work with a collection of about 40K news articles from CNN. More details about the dataset are available on Kaggle at:

https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning

You can download the dataset from there or from Moodle.

(a) As a first step, we need to import the dataset and transform it into a suitable representation.

  (i) Import the dataset into a pandas DataFrame and transform the column `ArticleText` using the `TfidfVectorizer` with its default settings.

  (ii) To get an idea for a reasonable number of clusters to aim for, determine how often each value of the two columns `Category` and `Section` occurs in the dataset.

(b) Now, we apply **k-Means** to the dataset.

  (i) Compute two clusterings using k-Means for values of $k \in \{6, 37\}$.

  (ii) For both clusterings, determine for each cluster the data point that is closest to the centroid. Output the text of these articles as a description of the clusters.

  (iii) Visualize the two determined clusterings in two-dimensional scatter plots. For each cluster, data points should be drawn in a separate color. To map the data points onto two dimensions, please make use of t-SNE (available in `sklearn.manifold.TSNE`) as a dimensionality reduction technique.

(c) Now, we apply **DBScan** to the dataset.

(i) Determine two choices of the hyperparameters ϵ and $minPoints$, so that the resulting number of clusters is in $[5, 10]$ and $[30, 40]$, respectively.

(ii) Visualize the two determined clusterings in two-dimensional scatter plots. For each cluster, data points should be drawn in a separate color. To map the data points onto two dimensions, please make use of t-SNE (available in `sklearn.manifold.TSNE`) as a dimensionality reduction technique.

(d) Finally, we want to quantitatively evaluate the four clusterings computed in (b) and (c).

(i) For each of the four clusterings, determine its Homogeneity score (as implemented in `sklearn.metrics.homogeneity_score`). Homogeneity is another external quality measure. For the smaller (larger) clusterings use `Section` and `Category` as a ground-truth labeling (idealized clustering).

(ii) For each of the four clusterings, determine its Silhouette score (as implemented in `sklearn.metrics.silhouette_score`). Silhouette is another internal quality measure.