TU 257 – Fundamentals of Data Science

Data Analytics

Lab 10 – Clustering Data

Brendan Tierney

Agenda

- Exercise 1 – Clustering Data using K-means
  - Part 1 – Examples/Demo Notebook 1
  - Part 2 – Write code to expand this Notebook by answering Exercises/Question throughout
- Exercise 2 – Density Based Clustering
  - Part 1 – Examples/Demo Notebook 2
- Exercise 3 – Select a Dataset and do some Clustering (Optional)

Exercise 1 - Clustering Data using K-means

Demo Notebook 1 – Clustering Data

- Part 1 – Examples/Demo Notebook 1
  - Download the Demo Notebook 1
  - Download the Dataset needed for Demo Notebook 1

  - Run all cells
  - Examine the output generated
  - Follow what is happening from cell-to-cell
  - Add extra annotations/descriptions based on your understanding
    - Enrich the descriptions to make them more meaningful for you

Demo Notebook 1 – Clustering Data

- Part 2 – Write code to expand this Notebook by answering Exercises/Question through
  - NB: Complete the previous Part 1 before commencing this following
  - Look through Demo Notebook 1. There are Questions and Exercises throughout
  - Write code or modify the existing code to answer these Questions and Exercises
    - Consider adding a new cell to contain the new or modified code

Exercise 2 – Density Based Clustering

Demo Notebook 2 – Density Based Clustering

- Download the Demo Notebook 2

- Run all cells
- Examine the output generated
- Follow what is happening from cell-to-cell
- Add extra annotations/descriptions based on your understanding
    - Enrich the descriptions to make them more meaningful for you

- Compare differences in outputs generated by DBScan and K-means
- Add addition code to use K-means++ algorithm
- Compare outputs from using K-means++ to previous results

Exercise 3 – Select a Dataset and do some Clustering (Optional)

Exercise 3 – New Dataset - Optional

- Select a suitable Dataset for clustering
- Using the sample code from the Demo Notebooks, build your own Clustering solutions
- Dataset Repositories
  - Check out the following dataset repositories
    - https://oralytics.com/2019/04/18/data-sets-for-analytics/
  - UCI Data Set Repository https://archive.ics.uci.edu/ml/datasets.php
    - Search the list of datasets for Clustering
    - Pick one that is of interest to you.

Any Questions ?

What Now/Next ?

Complete **all** Lab Exercises before Next Week

Pick another dataset and complete same/similar tasks with it