

Introduction to Data Science and AI: LAB 4

Tung Kieu

May 22, 2023

1. Requirements

- 1.1 **Integrated Development Environment (IDE):** PyCharm.
- 1.2 **Version Control:** Git and TortoiseGit.
- 1.3 **Compiler & Interpreter:** Python 3 (WinPython on Windows or Anaconda on Linux).
- 1.4 **Additional Libraries:** Pandas, NumPy, SciPy, Matplotlib, Sklearn, (and PyTorch).
- 1.5 **Data Sets:** Iris and MNIST

2. Assignment

- 2.1 Load data set IRIS (<https://archive-beta.ics.uci.edu/dataset/53/iris>) by using NumPy.
- 2.2 Apply k -mean to dataset IRIS (using Sklearn) with $k = 3$.
- 2.3 Report the accuracy.
- 2.4 Apply k -mean to dataset IRIS (using Sklearn) with different k
- 2.5 Visualize the accuracy result with different k .
- 2.6 Visualize the Within-Cluster-Sum of Squared Errors (WSS) result with different k .
- 2.7 (BONUS) Apply k -mean to dataset IRIS (using your writing code) with $k = 3$.
- 2.8 (BONUS) Apply k -mean to dataset IRIS (using your writing code) with different k
- 2.9 (BONUS) Visualize the accuracy result with different k .
- 2.10 (BONUS) Visualize the Within-Cluster-Sum of Squared Errors (WSS) result with different k .

- 2.11 Describe how to choose a good k .
- 2.12 Apply DBSCAN to dataset IRIS (using Sklearn).
- 2.13 Report the accuracy.
- 2.14 Apply DBSCAN to dataset IRIS (using Sklearn) with different eps and $min_samples$.
- 2.15 Visualize the accuracy result with different eps and $min_samples$.
- 2.16 Visualize the Within-Cluster-Sum of Squared Errors (WSS) result with different eps and $min_samples$.
- 2.17 Compare k -mean and DBSCAN on dataset circles.
- 2.18 Compare k -mean and DBSCAN on dataset moons.
- 2.19 Compare k -mean and DBSCAN on dataset blobs.

3. Submission

- 3.1 Write the report in LaTeX.
- 3.2 The report should have code, results, chart figures, and explanation.
- 3.3 The report should be submitted in one week after the lab date.