S k-means version 1

June 1, 2022

Đoàn Quang Văn

University of Tsukuba

Contact: doan.van.gb(at)u.tsukuba.ac.jp

This document is to explain the scripts reposited in the GitHub S k-means repository. These include python codes of k-means programs, wrap-up script of three test runs, and scripts for plotting a) Similarity-Distributions, b) Silhouette analysis results, c) Clustering uncertainty degree evaluations.

# Main program

|  |  |
| --- | --- |
| Scripts | Descriptions |
| kmean.py | * main program containing functions of the k-means algorithm   ﻿K\_means(***X,k,sim, ini****)*, where ***X*** is a set of input vectors, ***k*** is number of clusters. ***sim*** is classification scheme (one can select “ssim”, “str”, “ed”, “md”, which mean structural similarity, Pearson correlation coefficient, Euclidean distance, or Manhattan distance). ***ini*** is initialization scheme. One can select “rand” for randomized initialization, or “pp” for k-means ++ initialization scheme   * No furthermore libraries (other than default ones in anaconda framework) is needed to run the program. Seaborn lib is needed for some plot. * This script is self-standing. One can run it directly from terminal $ python kmean.py to generate some demonstration result. * Chart, scatter chart    Description automatically generatedIn demonstration runs, k-means is applied for randomly distributed 2-dimensional data. One can select preferable setting, by modifying some lines (for k, sim, etc.) * The clustering result is saved in demo/run\_?/ ﻿cluster.csv, and plotting are also automatically generated and saved in the same directory. For example, the scatter plot of the input data (right plot).   Also diagnosing results such as the Similarity-Distribution, Similarity-Matrix of the input vectors are also automatically generated (below)   * Chart, histogram    Description automatically generatedDiagram, schematic    Description automatically generated with medium confidenceClustering results are plotted automatically together with Silhouette analysis   Chart  Description automatically generatedChart, scatter chart  Description automatically generated   * Chart, bar chart, histogram    Description automatically generatedIn the case of multiple runs, one can assess the uncertainty/consensus of clustering results using Clustering Uncertainty Evaluation framework proposed together with S k-means. This analysis is also automatically conducted and the plot is generated ( right figure) |
| main\_kmean.py | * script to wrap up kmean.py for three test problems. With several number of loops commands, i.e., 3 (tests, ﻿'TC\_ll', 'AM\_t\_1950\_y', 'SLP\_DJF') x 4 (classification schemes, ﻿'ssim', 'ed', 'md', 'str') x 11 (number of cluster k ) x 10 (initializations) = 1320 runs are conducted. * Output data are generated and save into directory specified in the script, for each run the Silhouette analysis is also automatically conducted. The Silhouette score, and figure are automatically saved in the output directory (explained latter in Output section) |

# Input data

In this study, three demonstration problems are introduced.

<https://drive.google.com/drive/folders/13IGgRDKf5yhb9535_mWKvqpr0-ZTxM-o?usp=sharing>

|  |  |
| --- | --- |
| Script | Descriptions |
| SLP\_DJF.nc | main program of k-means algorithm, with four options of classification, i.e., S-SIM, COR, ED, MD  self-standing, one can run the rest directly, i.e., > python kmean.py  Diagram  Description automatically generated |
| AM\_t\_1950\_y.nc | script to wrap up kmean.py for three test problems. |
| TC\_ll.nc |  |

# Output data

<https://drive.google.com/drive/folders/1k8zRelePylyn2F_JQshrMSeAH5FV_Hdd?usp=sharing>

../output\_20220318/AM\_t\_1950\_y/00/n02/rand

../output\_20220318/AM\_t\_1950\_y/ed\_sim\_btw.nc

# Plotting

Scripts for plotting

1. Plot weather map from simulation WP
2. Plot climate change from CC
3. Plot tropical cyclone from TC

|  |  |
| --- | --- |
|  | Description |
| **plot\_fig02\_S-Ds.py** | Plot fig 02 in the manuscript |
| **plot\_fig03\_WP.py** |  |
| **plot\_fig04\_cc.py** |  |
| **plot\_fig05\_tc.py** |  |
| **plot\_fig0607\_gen\_res.py** |  |
| **plot\_chord.py** **plot\_fig08-10\_CUD.py** |  |
| **write\_docx.py** |  |

<https://drive.google.com/drive/folders/1BRft5aHLDTIZ7VGf6cCGBdwAJCFkWrhZ?usp=sharing>