Kmeans

April 20, 2020

1 problem definition

Kmeans is an algorithm unsupervised clustering, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

add subsection on : use cases of this algorithm + image ; the algorithm flow ; the complexity ; how paralelizable it is

2 serial implementation

This section concerns the `Kmeans_serial.py` python3 file.

Pre-requisites:

- numpy
- matplotlib
- scikitlearn

The provided code proposes a simple implementation of the Kmeans algorithm, as well as showcases the (optimized) version implemented in the *scikitlearn* library.

When launched, the code will :

- generate some random two-dimensional points from 5 populations with different means and variances.
- apply our Kmeans algorithm on the generated data, with `k=5`, and report execution time.

- plot the result of this clustering as a scatter plot where symbols denote the real population assignment, and colors correspond to the Kmeans' cluster assignment.
- perform the Kmeans clustering using scikitlearn implementation and report its execution time.