

Release of Code and Data for Angle K-Means



1 CODE REPOSITORY

```
anglekmeans/
├── demo.py          # demonstration script
├── pyproject.toml
├── README.md
├── setup.py
├── demo_data/
│   ├── RaFD_k50_centroids.mat # sample dataset
│   └── RaFD.mat              # sample centroid file
└── src/
    ├── anglekmeans/
    │   ├── __init__.py
    │   ├── mykm.py
    │   ├── gemm_mask_float.c
    │   └── akmc.c ...
```

Fig. 1: Key Directory Structure of the Angle k -means Implementation

1.1 Angle K-Means

This repository provides a demonstration script (`demo.py`), a sample dataset (`RaFD.mat`), and its centroid file at $k = 50$ (`RaFD_k50_centroids.mat`).

1.2 Environment Setup

All codes are implemented in C/C++, and python interfaces are provided. We run them on the Ubuntu 24.04.2 LTS machine with 64 GB main memory, 5.10 GHz i5-13600 CPU.

- 1) Create a new Anaconda environment named `py12` using the command:

```
(base) $ conda create -n py12 python=3.12
```

- 2) Activate the environment using the command:

```
(base) $ conda activate py12
```

After successful activation, the command line prompt will show something like:

```
(py12) $
```

1.3 Preparation & Installation

- 1) Firstly, the necessary libraries can be installed using the following command:

```
(py12) $ pip install numpy scipy cython scikit-learn; conda install -c conda-forge mkl-devel
```

- 2) Access the anglekmeans directory:

```
(py12) $ cd /path/to/anglekmeans
```

- 3) Finally, compile anglekmeans:

```
(py12) $ python setup.py build_ext --inplace
```

1.4 Running the Demo

After completing the preparation and installation setup, execute with:

```
(py12) $ python demo.py
```

1.5 Expected Output

For the dataset `RaFD.mat` and its centroid file `RaFD_k50_centroids.mat`, the expected output results are shown in the figure below. These results correspond to those presented in Table 2 of the main text for RaFD with $k=50$.

```
seed = 0

Data Info
n = 8040
d = 256
k = 50

Lloyd
Iteration = 8
Per-iteration average of distance computation = 4.0e+05
acc = 4.5e-01

Angle
Iteration = 8
Per-iteration average of distance computation = 1.7e+05
acc = 4.5e-01
Label difference = 0
```

Fig. 2: Expected Output After Running the Demo

2 DATA INTRODUCTION

2.1 Data Files

`anglekmeans` provides a sample facial data file, `RaFD.mat`, where the number of data points (n) is 8040, the data dimensionality (d) is 256, and the number of clusters (k) is 67.

2.2 Centroid Files

Since k-means involves randomness, for a fair comparison we used a fixed set of initial centroids in experiments. We provide a centroid file, `RaFD_k50_centroids.mat`. This file is the centroid file for the `RaFD.mat` dataset generated with $k=50$ clusters. The centroids are determined by `kmeans++(seed=0)`. The following content from the centroid file is readable.

- *centroid* is the cluster centroids (shape: $[k, d]$)
- *d* is the dimensionality of the data points
- *k* is the number of clusters