

COMS 4030A

Adaptive Computation and Machine Learning

LAB EXERCISE 5

- (1) Code up the k -MEANS ALGORITHM.
- (2) Code up the ONLINE k -MEANS ALGORITHM.
- (3) Code up the HIERARCHICAL CLUSTERING ALGORITHM.

You can create your own test datasets by sampling from normal distributions.

To create a test dataset of points in \mathbb{R}^2 , first choose some points $(a_1, b_1), \dots, (a_\ell, b_\ell)$ in \mathbb{R}^2 ; these will be the means of distributions.

For the first point (a_1, b_1) , choose a standard deviation σ .

Randomly sample 2 points in the normal distribution $\mathcal{N}(0, \sigma)$, say c, d .

Create a new point $(a_1 + c, b_1 + d)$ in \mathbb{R}^2 .

This point goes into your dataset, and you can repeat this a number of times using the same mean (a_1, b_1) to get a number of datapoints.

Repeat the above process with the other means $(a_2, b_2), \dots, (a_\ell, b_\ell)$; you can use the same standard deviation, or a different standard deviation for each mean.

(You can create test datasets in other \mathbb{R}^m 's similarly.)

The set of points you create should consist of ℓ clusters of points distributed around the means. When testing your algorithms on this dataset, if you set $k = \ell$, the final cluster centres produced by the k -means algorithms should correspond to the means of the distributions.

You can then try different values for k to see what happens.

For the k -means algorithms, after every loop over the dataset, calculate the sum-of-squares error and see if it decreases during training, and whether it converges or not.

For the Hierarchical Clustering, you can try different proximity metrics.

You don't need to submit anything for this lab exercise, but Assignment 2 will be based on it.