Title- Clustering based on a EM algorithm and K-Means

Apply EM Algorithm to cluster a set of data stored in a .CSV file. Use the same data set for clustering using K-Means algorithm. Compare the results of the these two algorithm the quality of clustering.

What is EM Algorithm

The Expectation- Maximization algorithm is an approach for performing maximum likelihood estimation in the presence in the presence of latent variables. It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence.

M-STEP

E -STEP

Gaussian Mixture Model

Gaussian Mixture Model are probabilistic models and use the soft clustering approach for disturbing the points in different clusters.

Gaussian Mixture Model (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distribution represent a cluster. Hence a Gaussian mixture model tends to group the data points belonging to a single distribution together.

E-M ALGORITHM

1. Given a set of incomplete data, consider asset of starting parameters.
2. E-Step: Using the observed available data of the dataset, estimate the values of the missing data.
3. M-Step: complete the data generated after the expectation step is used in order to update the parameters.
4. Repeat the step 2-3 until the convergences.

K-MEANS ALGORITHM

K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups where each data point belongs to only one group.

It tries to make the intra cluster data points as similar as possible while also keeping the clusters as different as possible.

It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum.

K-Means Algorithm

1. Specify the number k of clusters to assign.
2. Randomly initialize k centroids
3. Repeat
4. Expectation: Assign each point to its closest centroid
5. Maximization: Compute the new centroid of each cluster
6. Until The centroid positions do not change

IRIS DATASET

This datasets consist of 3 different types of irises( setosa, versi colour, and virginica) petal and sepal length, stored in a 50\*4 numpy.ndarray

The rows being the sample and the columns being, Sepal length, sepal width, petal length and petal width

The IRIS datasets is a classic and very easy multi class classification datasets.

|  |  |
| --- | --- |
| Classes | 3 |
| Samples per classes | 50 |
| Samples total | 150 |
| Dimensionality | 4 |
| Features | Real,positive |