Advanced Machine Learning – Practicum 1 – K-means Algorithm

Topics covered: K-Means Clustering Algorithm, Clustering Objective Function, Evaluating Clustering Algorithm

Deliverables:

Your submission for this assignment should be an archive of three files:

1. ***kmeans.py***

   This file should contain the following functions below:

Function name	Input	type	Output	type
`loadData`	Filename	String	X: dataset	array
`errCompute`	X: dataset, 2 features + 1 ClusterID	Array with shape (16259,3)	The value of objective function for the clustering	float
	M: mean value for each cluster	Array with shape (K,2)		
`calcMeans`	X: dataset, 2 features + 1 Cluster ID	Array with shape (16259,3)	M: the updated mean value for each cluster	Array with shape (K,2)
	M: Current means	Array with shape (K, 2)		
`Group`	X: dataset for training, 2 features + 1 ClusterID	Array with shape (16259,3)	X with updated clusterID. Assign each object into its closest cluster	Array with shape (16259,3)
	M: Current means	Array with shape (K,2)		

2. **<username>.pdf** should follow the outline described in *Section 3: Tasks* and you should put in all the required materials in Section 4.

You should zip the two files called **<username>_4.zip**. Also, add one more folder named **iniMeans** where you will store your initial means. This will be used to run your code with the exact initial-value and test result that you have made in your implementation. Failure to follow conventions will result in penalties in marks.

Objectives:

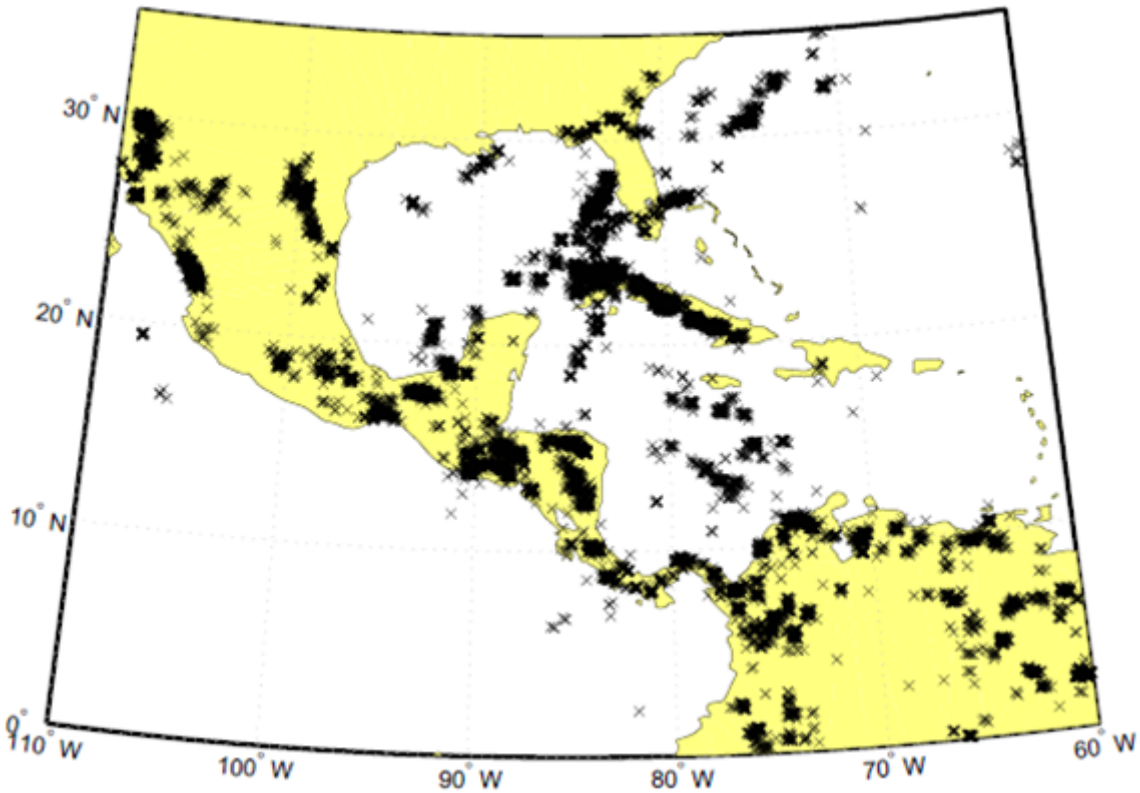
- To implement K-means algorithm to learn the clustering partitioning method from the training data.
- To apply K-means group dots representing lightning strokes.
- To get familiarized on determining the parameter K for K-means.

1. Dataset

Lightning networks detect individual lightning discharge events all over the world. For each lightning event, the data is stored as (time, amount of discharge, latitude, and longitude). It shows that when and where, there is a lightening stroke with certain amount of discharge. Clustering helps the meteorologists to group the lightning strokes basing on the position.

2010825.txt records the places of the lightning strokes during 10:00-23:59pm on the day 25/08/2010.

Below is the plot of the data.



The dataset in the file only contains two features for clustering:

1. X: x-position in coordinate system, which is converted from latitude.
2. Y: y-position in coordinate system, which is converted from longitude.

NOTE: a **test.txt** with 15 objects is also provided to help you verify your code. I suggest implement and test your code on test.txt first and then apply the code on lightening data.

2. Documentation

For documentation, you need to write a report on this assignment by following instructions in Tasks. You should submit ONLY the “pdf” of the document and NO other file formats will be accepted. A template in .doc format can be found in the assignment page.

3. Tasks

a. Write a function `loadData(filename)` to load lightening data into an array `X`. After appending the third column for storing the `clusterID`. `X.shape` should be (16259,3).

b. Plot the data in the given dataset to visualize the input of the clustering. Paste your plot in Section 1 Lighting data of your report.

c. Construct the clustering objective function for K-Means. Briefly explain this equation. Write down them in Section 2 Objective Function of your report. Note: In this assignment, you will use Euclidean distance metric in the objective function.

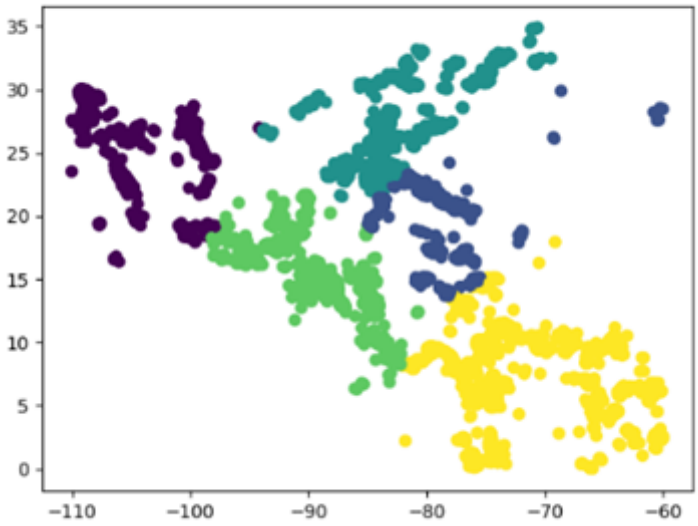
d. Use the objective function (from task c) to write a function `errCompute()` for evaluating the quality of clustering. This function takes in dataset `X` and means `M` (with shape (K,2)), returns the **mean** of the error `J`. Suppose that the cluster id for all objects in *X* are initialized as 0, and `M = np.array([[0,0]])`, run `errCompute(X,M)`, you should get a result of 86.62

e. Write down and briefly explain the method of assigning each object to a cluster in Section 3 assigning objects of your report.

f. Use the assigning method from task e to implement function `Group()`: to assign each object into a cluster basing on the current set of means `M`. This function takes in dataset `X`, `M` the current means. It returns the clustering result included in `X`. Suppose K=5, and the top 5 objects from *X* are chosen as initial means. i.e. `M=np.copy(X[0:5,0:X.shape[1]-1])`. After running `X=` Group(X,M)`, you should get` errCompute(X,M) = 9.73`

g. Describe the method of calculating new means basing on current clustering in Section 4 Updating Means of your report.

h. Use the method described in task g to implement the function calcMean(). This function takes in dataset *X*, *M* the current means and returns the updated *M*. K is still set as 5, repeatedly run *calcMeans*() and *Group*() until there is no changes in clustering result, you'll get *errCompute*(X,M)=4.01. The clustering result (k=5) shown below:



i. Run your k-means when K is chosen as 50 and 100 respectively and plot the clustering result. Compare the objective function values (i.e. value returned by `errCompute()`) for the two clustering results, and discuss which one is a better result. Write down you discussion in Section 5 Choosing K.

4. Rubrics

This assignment is graded over total of 100 points. The breakdown is as follows:

- task a and task b Data Load and Plot (10 points)
- task c: proper construction of K-Means objective function (10 points)
- task d: error function implementation (20 points)
- task e: objects assigning description (10 points)
- task f: group/assigning function implementation (20 points)
- task g: means calculation description (10 points)
- task h: means calculation implementation (10 points)
- task i: clustering result plotting & proper discussion on choosing K value (10 points)

Note: Submission requirements must be met i.e., reasonable comments, proper format of the report (Not copy and pasted from the assignment specifications!), detailed explanation (using necessary equations, tables, figures, charts, etc.), correct documentation and file formats for the submission, etc. If any of these requirements are not satisfactory, then zero marks for this assignment.