DLI Accelerated Data Science Teaching Kit

Module 15 Lab:
KMeans Clustering

OBJECTIVE

As an unsupervised learning method, KMeans finds k groups of data in a given dataset based on distances between data. The goal of this lab is to implement the KMeans algorithm for clustering data.

KMeans Algorithm:

1. Select k random points as initial centroids
2. Calculate distances between points in dataset and centroids based on Euclidean distance
3. Assign each point to the closest centroid
4. Find mean or each group of points and set it as new centroids
5. Check if centroids moved more than a predefined threshold. If no, repeat steps 2-5, if yes, algorithm is converged.

PREREQUISITES

Install Python packages below.

- **numpy** is the fundamental package for scientific computing with Python.
- **matplotlib** is a famous library to plot graphs in Python.
- **sklearn** is contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- **cv2** is a library to develop real-time computer vision applications.
- **random** implements pseudo-random number generators for various distributions.

INSTRUCTIONS

- Build the KMeans clustering algorithm
- Implement clustering with KMeans on 2D data
    - Building 2,000 samples with 2 features based on make_blobs() from sklearn library
    - Clustering 2,000 samples
    - Visualizing clustering results


NVIDIA
DEEP
LEARNING
INSTITUTE
Georgia
Tech
PRAIRIE VIEW
A&M UNIVERSITY