UECS3213 / UECS3453 Data Mining

SESSION: January 2019

Lab 9: k-Means Clustering Implementation Using scikit-learn

Introduction

Scikit-learn is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

In this tutorial, you will learn about the inner workings of the K-Means clustering algorithm with an interesting case study. K-Means falls under the category of centroid-based clustering. A centroid is a data point (imaginary or real) at the center of a cluster. In centroid-based clustering, clusters are represented by a central vector or a centroid. This centroid might not necessarily be a member of the dataset. Centroid-based clustering is an iterative algorithm in which the notion of similarity is derived by how close a data point is to the centroid of the cluster.

Objectives

At the end of this lab, you are expected to acquire the following:

a) The inner workings of the K-Means algorithm
b) A simple case study in Python scikit-learn
c) The disadvantages of K-Means

Instruction

1. Attempt the tutorial "Unsupervised Learning in Python" at the following link first at https://www.datacamp.com/courses/unsupervised-learning-in-python.

2. Then, attempt the "K-Means Clustering in Python with scikit-learn" at the following link: https://www.datacamp.com/community/tutorials/k-means-clustering-python.

3. Follow the step-by-step instructions in the tutorial.

Other Related References

- https://www.datasciencecentral.com/profiles/blogs/python-implementing-a-k-means-algorithm-with-sklearn

- https://stackabuse.com/k-means-clustering-with-scikit-learn/

The End