ENGINEERING DIVISION | NYU ABU DHABI

ENGR-UH 4560
Selected Topics in Information and Computational Systems

Machine Learning

Project 03 - K-means and Hierarchical Clustering

Introduction

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition ***n*** observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, Divisive and Agglomerative.

Divisive method

In this method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation.

Agglomerative method

In this method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters until there is only a single cluster left.

Requirements

- Implement your own K-means clustering module.
- Test your K-means module on example dataset (*ex7data2.mat*, *ex6data3.mat*).
    - Plot the output clustered results.
- Image compression with K-mean clustering module (*bird_small.png*).
    - Compress the example image via your own K-means module.
    - Plot the output image.

Optional task: (50 points bonus)

- Implement a hierarchical clustering model (*Mall_Customers.csv*).
    - Use *'Ward*' distance matrix for this dendrogram which is also known as the incremental algorithm.

      $$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}$$

      u is the newly joined cluster consisting of clusters s and t
      v is an unused cluster in the forest

T = |v|+|s|+|t|
|*| is the cardinality of its argument

- Build the model of Hierarchical Clustering.
- Plot the clusters and label customer types.
- Verify the optimal cluster number by your K-means clustering module on example dataset.

Deliverables

A zip file containing the following:

1. a working project (source code, makefiles if needed, etc)
2. a report for the detailed description of the project
    a. explain the main aspects of your code
    b. how to run your project
    c. plots and diagrams

Before submitting your project, please make sure to test your program on the given dataset.

Notes

Functions from Standard Python libraries (e.g. sklearn.cluster) are ***not allowed*** *in the task, you are encouraged to write your own module with similar performance.*
You may discuss the general concepts in this project with other students, but you must implement the program on your own. ***No sharing of code or report is allowed.*** *Violation of this policy can result in a grade penalty.*

Late submission will be accepted ***upon approval from the instructor****.*