IB120/201 - Lab 12

Clustering

Due Date: April 24, 2020

University of California, Berkeley *GSI: Naveed Ziari*

Clustering is an important tool in unsupervised learning because it helps biologists find meaningful relationships in multi-dimensional datasets. In this lab we will learn how to perform clustering in Python.

Background

K-Means

K-means clustering is a method of clustering n observations into k based on distance to the mean of the mean of the cluster. In more technical terms, it aims to produce a set of clusters that minimize the variance. The way the algorithm achieves that is by minimizing the sum-of-squares from the observation to the mean of the cluster:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Hierarchical Clustering

Another form of unsupervised clustering is hierarchical, whereby clusters of data are arranged by hierarchy. There are generally two algorithmic approaches: top-down and bottom-up. Top down algorithms assume one cluster in the beginning (at the top of the hierarchy) and recursively split the observations into separate clusters. Bottom-up approaches assume the opposite - each observation is in its own cluster (starting at the bottom of the heirarchy) and then joined together at each iteration. Hierarchical clustering draws analogy to phylogeny, whereby each taxonomic rank such as species, genus, family, order, etc. is a hierarchy. In this lab we will learn about two different types of bottom-up hierarchical clustering: centroid and UPGMA.

Questions

Please submit your assignment in the form of a iPython notebook similar to lecture. Each question has its own block.

1. Perform another type of clustering of your choosing on the `lab_12_data_01.xlsx` dataset and explain which type it is.
2. From the KMeans clustering demo in lab section, find the variance within each cluster. You will have to call upon the `labels_` attribute instead of `cluster_centers_` to assign a cluster index to each row in the data frame and then use the `.iloc` to extract the desired rows for calculation of the variance.
3. Read in the file `lab_12_data_02.xlsx` and make a new `Data.Frame` out of the columns `Annual Income (k$)` & `Spending Score (1-100)`.
   (a) Perform centroid hierarchical clustering on the dataset and visualize it in a dendogram.
   (b) Perform UPGMA hierarchical clustering on the dataset and visualize it in a dendogram.

   For the clustering, use the `linkage` and `dendogram` functions in `scipy.cluster.hierarchical`. read the documentation to see which parameter to adjust to get the desired algorithm

Use the `linkage` function from `scipy.cluster.hierarchy` to create the clusters `dendogram` function in `sklearn` to visualize the clusters.

4. From the lecture notes, please explain in your own words what the *curse of dimensionality* means and what could be done to overcome it when analyzing high-dimensional data.

5. Name a few types of distance metrics used in clustering algorithms.

6. Did we employ bottom-up or top-down hierarchical clustering methods in this lab?