Assignment 9 - Clustering and Web application.

Sebastian J. Bentley — cph-sb287

May 3, 2022

Link to GitHub: https://github.com/SebastianBentley/DataScienceAssignments/tree/main/assignment9

1 In Exercise 1 we used Hierarchical clustering algorithm.

1.1 Which type?

The bottom-up approach: **Aggromerative Hierarchical**

1.2 How many types of hierarchical clustering are you familiar with?

The bottom-up approach: **Aggromerative Hierarchical** Clustering and the top-down: **Divisible Hierrchical Clustering**

1.3 How do they differ?

"Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy." - https://en.wikipedia.org/wiki/Hierarchical_clustering

2 Train a clustering model with Mean Shift algorithm and freedom.csv data source file.

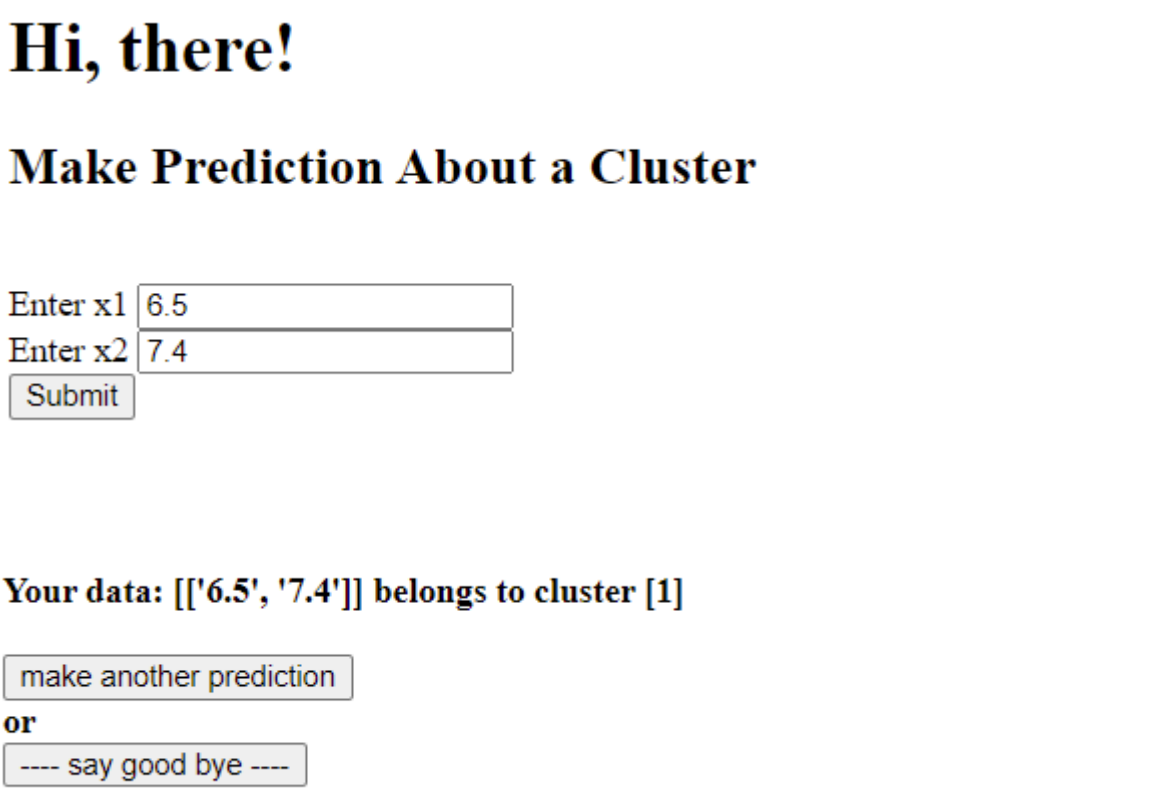
2.1 Store your model in a file

2.2 Create simple web application that can deploy and run the model as seen in class

2.3 Run the application for predicting the cluster of a data set

Solution for **2.1, 2.2, 2.3** can be found on GitHub, linked at the top of this document.

2.4 Take and attach a screen shot of your solution



3 Describe the difference between K-means and Mean Shift algorithms

3.1 In which occasions would you prefer to use the mean shift algorithm?

With inspiration from E12-1-Hierarchical.ipynb from the course exercises.

K-Means Clustering is simple to understand, Easily adapptable and efficient and works well on both small and large datasets. However, you need to know the optimal number of clusters in advance.

Hierarchical Clustering can find the optimal number of clusters from the model itself and dendograms are practical and easy to understand. However, it is not suitable for large datasets