CPSC 483 - Introduction to Machine Learning

Project 7, Fall 2020

due December 14 (Section 02) / December 17 (Section 01)

Last updated Tuesday November 24, 3:50 pm PST

In this project we will compare the results of using clustering algorithms to segment customers based on their purchasing data.

The project may be completed individually, or in a group of no more than three (3) people. All students on the team must be enrolled in the same section of the course.

Platforms

The platform requirements for this project are the same as for previous projects.

Libraries

You will need pandas to load the dataset, scikit-learn to run clustering algorithms, and SciPy to visualize hierarchical clustering with a dendrogram.

You may reuse code from the Jupyter notebooks accompanying the textbook and from the documentation for the libraries. All other code and the results of experiments should be your own.

Datasets

Download `groceries.csv`, a dataset consisting of weekly spending by families on several common grocery items. Spending is listed in units of dollars per person per week.

Experiments

Run the following experiments in a Jupyter notebook, performing each action in a code cell and answering each question in a Markdown cell.

1. Use `read_csv()` to load and examine the dataset.

2. Use `sklearn.cluster.KMeans` to cluster the dataset using the default parameters. Assign the `labels_` attribute determined by the algorithm to a new column in your DataFrame. How many clusters are there?

3. Use `pandas.DataFrame.groupby()` to group the data by cluster assignment, then use the `GroupBy` object to examine descriptive statistics such as minimum, maximum, and mean. Describe any differences you see between the clusters.

4. There's no particular reason to think that the default value of *K* is the correct one for this dataset. Let's switch to hierarchical clustering to see if we can visualize how the data clusters together.

   Plot a `scipy.cluster.hierarchy.dendrogram()` for the dataset. Single linkage may not be the best choice for the dataset, but it is relatively quick to compute and will help us make a decision on the number of clusters.

   Based on the dendrogram, how many clusters appear to be present in the dataset?

5. Repeat experiment *(2)* using the number of clusters you determined to be present in experiment *(4)*.

6. Now switch to `sklearn.cluster.AgglomerativeClustering()` and repeat for the same number of clusters. How do the label assignments compare for the two algorithms?

   Caution: don't include the cluster number assigned by *K*-Means as one of the features to be clustered, or it will skew your results.

7. Repeat experiment *(3)* with the clusters you obtained in experiment *(5)*. How would you describe the various types of customers? What can you determine about them based on the data? For each segment identified, list some other items they are and are not likely to buy in future shopping trips.

Submission

Submit your Jupyter `.ipynb` notebook file through Canvas before class on the due date. Your notebook should include the usual identifying information found in a `README.TXT` file.

If the assignment is completed by a team, only one submission is required. Be certain to identify the names of all students on your team at the top of the notebook.