Lab 4: Cluster Analysis

This lab is due at 11:59 PM ET on Sunday, October 28, 2018. Late submissions will be penalized as described in the syllabus.

You may work individually or in groups of two. If working in a group, only one student needs to submit the group's work via Carmen. As with previous labs, you'll do all your work in a Jupyter notebook. This lab is worth 100 points.

In this lab, you'll conduct a cluster analysis using the "Wholesale Customers" dataset.

You will use the following Scikit-Learn clustering classes:

- **`sklearn.cluster.KMeans`**
- **`sklearn.cluster.AgglomerativeClustering`**
- **`sklearn.cluster.DBSCAN`**

0. Get the data

Download the "Wholesale Customers" dataset from the UCI Machine Learning Repository. The dataset contains 440 rows and 8 columns.[1]

1. Preliminary data analysis (15 points)

Discuss any observed patterns or data quality issues, such outliers, missing values, and attribute correlations. Visualize those that are interesting or important. Include a dataset description and summary statistics, as always. Describe your observations.

2. Data preprocessing (5 points)

Determine whether any transformations or data preprocessing (e.g. feature subset selection, scaling, feature creation) are needed and make the appropriate changes as you see fit. Describe and justify your decisions—i.e. what did you do and why?

You do not need to split your data into training and test datasets for this lab.

3. Find "natural" clusters (70 points)

Without any assumptions about K, try to find an optimal clustering using all three clustering estimators (K-means, hierarchical, and DBSCAN) as well as the evaluation measures and methods (cohesion, separation, silhouette score, elbow method, etc.) we discussed in class. Discuss your findings and provide appropriate visualizations. Describe the estimator that gave the best results.

[1] archive.ics.uci.edu/ml/datasets/Wholesale+customers

4. Exceptional work (10 points)

You may complete this part as a group. Choose ONE of the following:

- Assume K=2 and try to find an optimal clustering using any or all of the three techniques (K-means, hierarchical, and DBSCAN). Evaluate your results in terms of accuracy with respect to the "Channel" attribute. (You must not include the "Channel" attribute in your cluster computations.) Discuss your findings and provide appropriate visualizations.
- Assume K=3 and try to find an optimal clustering using any or all of the three techniques (K-means, hierarchical, and DBSCAN). Evaluate your results in terms of accuracy with respect to the "Region" attribute. (You must not include the "Region" attribute in your cluster computations.) Discuss your findings and provide appropriate visualizations.
- Use principal component analysis (PCA) or Isometric mapping (Isomap) to reduce the dimensionality of the dataset, then see if you can improve upon your best clustering results from part 3. Discuss your findings and provide appropriate visualizations.

5. Submit your work

Be sure to report any problems or bugs you could not resolve. Save your notebook using descriptive filename, such as **`lab4_burkhardt_5.ipynb`**. You may split your work into multiple notebooks if you wish but be sure to provide consistent and meaningful filenames. Submit your notebook(s) via Carmen, under Lab 4. Do not submit any data files.