CAP 6610 - Machine Learning (Spring 2022)
Assignment 3 - Due: Friday, 4/1/2022 11:59 pm

This assignment is a programming assignment, where you will implement a machine learning system to solve a clustering problem.

You are given the **Online Shoppers Intention** dataset and you are to cluster these data to provide helpful insights. The details of these two datasets and the questions you will try to answer are as follows.

1 The Data

For this second part, you are going to consider the Online Shoppers Intention dataset (provided at UCI Machine Learning Repository). This dataset too is accessible on our Canvas course site.

This dataset has 18 columns and 12,330 rows. The 18 attributes include 4 categorical and 14 numerical attributes. The last attribute "Revenue" is the class label: "FALSE" means not ending up shopping, and "TRUE" means ending up shopping. The meaning of the other attributes are the following.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

The “Bounce Rate”, “Exit Rate” and “Page Value” features represent the metrics measured by “Google Analytics” for each page in the e-commerce site. The value of “Bounce Rate” feature for a web page refers to the percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session.

The value of “Exit Rate” feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. The “Page Value” feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

The “Special Day” feature indicates the closeness of the site visiting time to a specific special day (e.g. Mothers Day, Valentine’s Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentins day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

As far as the categorical features, **feature engineering** is needed for the clustering methods we use for this assignment. From the dataset, you will see that Weekend and Revenue are binary, Month has 10 unique values, and VisitorType has 3. To numericalize them, Weekend and Revenue will be turned into 0 for FALSE and 1 for TRUE. You will use **Mean Encoding** for Month and VisitorType. Details and helpful examples on Mean Encoding can be found at this url: `https://www.geeksforgeeks.org/mean-encoding-machine-learning/`

2 The Task

You are to explore the following clustering models with $k = 4$ to provide insight of the dataset and report the comparison of their performances.

1. K-means [1]
2. Complete-Linkage Agglomerative nesting [2]

When clustering, do **NOT** consider the last attribute “Revenue.”

3 Performance Measures

Let us take the last attribute “Revenue” as the label and let us denote by $\mathcal{C} = \{C_1, C_2\}$ the two clusters it gives. Similarly, we denote by $\mathcal{C}^* = \{C_1^*, \ldots, C_4^*\}$ the clusters the model generates. Let us define $\lambda_i \in \{1, 2\}$ to be the cluster label of example $\boldsymbol{x}_i$ in clustering $\mathcal{C}$, and $\lambda_i^* \in \{1, 2, 3, 4\}$ cluster label of $\boldsymbol{x}_i$ in clustering $\mathcal{C}^*$.

We now define two sets S and D:

- $S = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) | i < j, \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*\}$
- $D = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) | i < j, \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*\}$

[1] sklearn.cluster.KMeans

[2] sklearn.cluster.AgglomerativeClustering

Intuitively, S is the set of all example pairs that are labeled the same in $\mathcal{C}$ and that are put in the same cluster in $\mathcal{C}^*$, and D is the set of all example pairs that are labeled differently in $\mathcal{C}$ and that are put in different clusters in $\mathcal{C}^*$.

The **Rand Index** (RI) takes S and D and computes $RI = \frac{2(|S|+|D|)}{m(m-1)}$, where m is the number of examples. RI is a value in the unit interval $[0, 1]$, the bigger the better the model. Thereafter, you will use RI to compare k-means and agglomerative nesting.

4 Requirements

1. Your programs should be in Python 3 and you are free to use any Python package to help you develop your programs.
2. There will not be training or testing, and you will use the whole dataset for clustering.
3. Report which model, K-means or Complete-Linkage AGNES, is better considering their RI scores.

5 Deliverables

Zip the following to [your-last-name]_Assignment3.zip and submit to Canvas.

1. A directory that contains all your Python programs that are may be .py or .ipynb.
2. A README file that contains instructions to run your Python programs.
3. A TXT report that describes the clustering results of the two models and compare their RI scores to report which one is better.