INTRODUCTIO TO DATA SCIENCE

3rd Homework Assignment

Due on: May 15, 2024

The goal of this homework assignment is to explore the **KMeans** algorithm using the given dataset **airlines.csv**. Throughout this assignment, you will perform various tasks including data description, data preprocessing, exploratory data analysis, and determining the optimal number of clusters using **KMeans**.

Tasks

1. *Data Description*

The provided raw data is in the airlines.csv file.

The description of the raw data is as follows:

id: Unique ID
balance: Number of miles eligible for award travel
qual_mile: Number of miles counted as qualifying for Topflight status.
cc1_miles: Number of miles earned with freq. flyer credit card in the past 12 months:
cc2_miles: Number of miles earned with Rewards credit card in the past 12 months:
cc3_miles: Number of miles earned with Small Business credit card in the past 12 months:

- **1**: under 5,000
- **2**: 5,000 - 10,000
- **3**: 10,001 - 25,000
- **4**: 25,001 - 50,000
- **5**: over 50,000

bonus_miles: Number of miles earned from non-flight bonus transactions in the past 12 months.
bonus_trans: Number of non-flight bonus transactions in the past 12 months.
flight_miles_12mo: Number of flight miles in the past 12 months.
flight_trans_12: Number of flight transactions in the past 12 months.
days_since_enrolled: Number of days since enrolled in flier program.
award: whether that person had an award flight (free flight) or not.

2. Check for Missing Values

Perform data preprocessing to check for any missing values in the dataset.

3. Analyze Features

Create histograms to understand the distribution of different features in the dataset.

4. Calculate Percentage of Customers with/without Award

Find the percentage of customers who do not have an award flight and those who do have an award flight.

5. Correlation Analysis

- Find which feature is correlated with the balance feature.
- Draw a correlation heatmap to visualize the correlations among different features.

6. Plotting

Plot the relationship between frequent flying bonuses and non-flight bonus transactions.

7. Determining Optimal Number of Clusters

- Apply MinMaxScaler to normalize the data.
- Use the Elbow Method and Silhouette Score to find the ideal number of clusters for KMeans algorithm.

Submission:

Submit your Python code or Jupyter Notebook in a file named your-student-id.py (e.g. 490606-hw3.py or 490606-hw3.ipynb) through Itslearning. Please upload the Python/Notebook file only.

IMPORTANT

➢ Academic dishonesty, including but not limited to cheating, plagiarism, and collaboration, is unacceptable and subject to disciplinary action. Any student found guilty will have a grade of F. Assignments are due in class on the due date. Late assignments will generally not be accepted. Any exception must be approved. Approved late assignments are subject to a grade penalty.