Mobile User Segmentation

Unsupervised Learning

This exercise is part of the graded case study on Unsupervised Learning.

Domain

Mobile

Business Context

A key challenge for Mobile App businesses is to analyze the trend in the market to increase their sales/usage.

We have access to the user's demographic characteristics, geolocation, and mobile device properties. This grouping can be done by applying different criteria like user’s data, their age group, phone brand compatibility and so on.

The machine learning clustering algorithms can provide an analytical method to cluster user segments with similar interests/habits.This will help App/mobile providers better understand and interact with their subscribers.

Objective

We will be clustering the users into groups by selected features that significantly distinguish different brands from each other and understand which factors are responsible for making the clusters

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Dataset description

events.csv - Event data has an event id, location detail (lat/long), and timestamp, when the user is using an app on his device

gender_age.csv - details of users age & gender

phone_device.csv - Device ids, brand, and models name. here the brands names are in Chinese, you can convert it in english using google for better understanding but we will not do it here.

Evaluation:

1. Preprocessing the data (5 points)
    a. Import required libraries and read all the 3 csvs.
    b. Analyze each csv by checking a few samples, drop duplicates.
    c. Check distribution of important features like gender, age group, etc.
    d. Merge the 3 csv into a single dataframe.

2. Exploratory Data Analysis (10 points)
    a. Check dimensions of the dataframe in terms of rows and columns and study few of the variables
    b. Check the data types
    c. Check the frequency and distribution of the relevant features
    d. Convert string features (phone_brand, device_model and gender) into categories and make them numerical
    e. Study summary statistics and mention your findings
    f. Check for missing values and impute/drop missing values if any
    g. Drop irrelevant columns like 'timestamp','event_id','device_id'
    h. Standardize the data to bring relevant features into a scale

3. Build a clustering algorithm for clustering mobile users. Kindly follow the below steps: (10 points) [Hint - you can try both k-means and hierarchical clustering]

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

a. Sample only 20000 data points from the standardized dataframe(you can take lesser/higher number of features depending on your computer)
b. Apply K-means or hierarchical clustering to the standardized dataframe
c. Mention the hyperparameters that perform the best (for eg: if you're using K-means , you can find K value that gives best silhouette score)
d. Evaluate the clustering algorithm you've used

4. Cluster Profiling: (10 points)
    a. Add a cluster label column created using the clustering algorithm in your original dataframe
    b. Compute the statistical summary for observations in each cluster(Check mean, sd, freq, modes, min, max, range..all basic central tendency numbers for each cluster(hint: dataframe[dataframe.cluster==0].describe() for 0th cluster, similarly do for other clusters))
    c. Perform bivariate analysis between cluster labels and other features

5. Do dimensionality reduction using PCA (5 points)
6. Apply k means/hierarchical clustering(depending on which you've implemented) on the PCA transformed data (5 points)
7. Mention your comments and findings from clustering profiling (5 points)

Optional:

1. Try KPrototypes algorithm to cluster the data
    a. https://medium.com/@guruprasad0o_o0/notes-on-k-prototype-for-clustering-mixed-typed-data-e80eb526b226
    b. https://medium.com/datadriveninvestor/k-prototype-in-clustering-mixed-attributes-e6907db91914

2. Try kmodes library to cluster the data
    a. https://pypi.org/project/kmodes/
    b. https://medium.com/@davidmasse8/unsupervised-learning-for-categorical-data-dd7e497033ae

Food for thought:

Does applying PCA give a better result in comparison to earlier?
Can you apply any other algorithms to create clusters of data?

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

How clustering can be helpful for your analysis?
What can you infer about the properties of users of different clusters formed in this project?

Learning Outcomes

- PCA
- k-means clustering
- Scaling
- Silhouette Coefficient

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.