ESE 5410: Machine Learning for Data Science

Chapter 9 Project

In addition to answering the **questions** on Canvas, also submit your Project Notebook in the default location within Codio. Please follow the instructions below (e.g., setting random state values) to ensure that your answers match the solutions.

You currently work for a company that offers a subscription-based music streaming service. After an unsuccessful round of marketing efforts with low return on investments, your company is exploring other strategies to promote its services. Currently, its advertising strategy has been to bid on a set of keywords that someone on the marketing team put together, but the team has realized that they could improve on this strategy by refining the keywords based on different customer personas. Therefore, the company has asked you to help it identify a set of user segments so that the ads can be more customized.

For this task, you have been provided with the Music data which contains sociodemographic and music preference data on 4,914 users and your goal is to identify clusters of users based on this information.

- Age : Age of the user
- Gender : Gender of the user
- Employment status : Employment status of the user
- Annual income : Annual income of the user in USD
- Usage per month : Average usage per month of the user measured in minutes
- Top genre : The genre of music that is most streamed by the user
- Num of days active : The number of days in the last 365 days that the user used the service

1 Part A: Data Cleaning

Throughout this course, you have mostly been provided with relatively clean datasets but in the real world, data is usually a lot messier. This could be caused by data logging errors, bugs in the pipeline or even erroneous human input in surveys. Some common approaches of dealing with

these observations are by dropping them, replacing the erroneous values with the mean/median, winsorizing, etc. Here, we will just drop them.

1. Load the Music data and drop any rows with `NaN/null` values.
2. Plot bar charts and histograms of the data to visualize the distributions. Does anything seem unusual? **Answer on Canvas:**
   (a) **Are there more male or female users?**
   (b) **Which genre is the most popular among users?**
   (c) **How many users are students?**
3. Take a closer look at 'age' and 'num of days active'. Does anything look peculiar? **Answer on Canvas: How many rows are peculiar?** Drop these rows. *Hint*: Recall the values that age and number of days active can take on.

2 Part B

1. Use K-means clustering to find a set of user groups using the following features as inputs: 'age', 'annual income', 'usage per month' and 'number of days active'.
   (a) Standardize the data. *Hint*: `sklearn.preprocessing.scale` may be helpful.
   (b) Calculate the sum of squared distances of observations to their closest cluster center for $K \in [1, 10]$ and add these values to a list called `ssd`. Use the default hyperparameters and set `random_state=42`. **Answer on Canvas: What is the minimum number in `ssd`?** *Hint*: You may want to write a for-loop and refer to the `sklearn.cluster.KMeans` documentation.
   (c) Plot the values in `ssd` against the number of clusters, K for $K \in [1, 10]$.
   (d) Based on this plot, what is the best number of clusters to set using the Elbow Method? (More information on the Elbow Method can be found here.)
   (e) Assign this value to the variable `optimal_K` and **input your answer on Canvas**.
   (f) Retrain the K-means clustering algorithm with `n_clusters=optimal_K`. This will be your final model.

3 Part C

As a data scientist, it is important to be able to translate your findings to colleagues who may not have the same level of technical knowledge as you do; visualizations help a lot! Therefore, you have

decided to plot the clusters. However, since there are four dimensions that you want to plot – age, annual income, usage per month and number of days active – you will need to utilize dimensionality reduction techniques to be able to plot this on a 2-D plane.

1. Train another K-means model using the following features as inputs: 'age', 'annual income', 'usage per month' and 'number of days active'. Use `n_clusters=2` and `random_state=42`.

2. Find the first two principal components of the scaled data.
   Hint: Use `sklearn.decomposition.PCA` with `random_state=42`.

3. Plot a scatterplot with the 1^{st} principal component on the x-axis and the 2^{nd} principal component on the y-axis. Color each point by the cluster that it is in from (1). *Hint*: Check out `seaborn.FacetGrid`. **Answer on Canvas: Which plot is the most accurate?**