Final Project: Data Mining

November 15, 2024

1 BACKGROUND

There are two parts of the project: classification and clustering. You may work in a group of 2 members for this project. Each team/submission will submit a final **notebook report**. Data can be found under `http://pluto.hood.edu/~dong/datasets/`.

2 Part I

A new title, The Art History of Florence, is ready for release by The Charles Book Club (CBC). To design targeted marketing strategies, CBC has sent a test mailing to a random sample of 4,000 customers from its customer base. The customer responses have been collated with past purchase data.

2.1 DATASET

Use CBC_3200.csv and CBC_800.csv for this part. Each row in the spreadsheet corresponds to one market test customer. Each column is a variable with the header row giving the name of the variable. The variable names and descriptions are given below:

- Seq#: Sequence number in the partition
- ID#: Identification number in the full (unpartitioned) market test data set
- Gender: O=Male, 1=Female
- M: Monetary- Total money spent on books
- R: Recency- Months since last purchase
- F: Frequency - Total number of purchases
- FirstPurch: Months since first purchase
- ChildBks: Number of purchases from the category: Child books
- YouthBks: Number of purchases from the category: Youth books
- CookBks: Number of purchases from the category: Cookbooks
- DoItYBks: Number of purchases from the category Do It Yourself books
- RefBks: Number of purchases from the category: Reference books (Atlases, Encyclopedias, Dictionaries)
- ArtBks: Number of purchases from the category: Art books
- GeoBks: Number of purchases from the category: Geography books
- ItalCook: Number of purchases of book title: “Secrets of Italian Cooking.”
- ItalAtlas: Number of purchases of book title: “Historical Atlas of Italy.”
- ItalArt: Number of purchases of book title: “Italian Art.”
- Florence: =1 “The Art History of Florence” was bought, =0 if not

2.2 Project Goal

Which team or submission can most correctly predict whether a customer will buy The Art History of Florence? Use at least two performance metrics, one of which should be the area under the curve (AUC) score.

Training, Validation, and Testing In machine learning, the dataset is divided into three sets: the training set, used to train the model; the validation set, employed for hyperparameter tuning and performance evaluation during training; and the test set, reserved for a final unbiased assessment of the model's generalization to new data. The training set is the largest, teaching the model by exposing it to diverse examples. The validation set aids in preventing overfitting, guiding adjustments to hyperparameters, while the test set serves as an independent benchmark for evaluating the model's real-world performance, ensuring it hasn't memorized the training data but can generalize effectively. Use CBC_3200.csv for training and validating your models, and CBC_800.csv for testing only after a model is trained.

2.3 Requirements

- Describe any exploratory analysis performed. For each analysis, include why it is done, the findings, and whether or how it impacts later project stages.
- Describes any changes/pre-processing you made to the data set – for example, handling missing values, transforming variables, binning variables, handling class imbalance, or eliminating outliers, Elaborate why these operations are performed. Again, this is an open-ended question.
- Compare and contrast at least two data mining models. Make sure to provide a summary of model performance.

3 Part II

This project aims to use unsupervised clustering techniques on structured electronic health record (EHR) data, including patient demographics, diagnoses, and lab results, to identify distinct subgroups of patients with obstructive sleep apnea (OSA) and heart failure (HF).

3.1 Dataset

Use osa.csv for part II. This dataset consists of more than 700 features. The features are encoded and scaled.

3.2 Project Goals

Which team or submission can better cluster patients into more than three groups? Use at least two performance metrics, one of which should be the Silhouette score.

In high-dimensional spaces, the distances between points "converge" because the relative difference between the distances of the nearest and farthest points becomes small. This makes it difficult for clustering algorithms to meaningfully differentiate between points based on distance, leading to

poorer performance and less distinguishable clusters. This article here describes one approach to address high dimensionality: the link.

3.3 Requirements

- Preprocess structured EHR data. Focus on dimension reduction.
- Apply unsupervised clustering algorithms to identify subgroups of patients with OSA and HF based on their demographics, diagnoses, and lab results. Use at least two clustering algorithms.
- Evaluate the clustering performance using appropriate metrics (e.g., silhouette score, Davies-Bouldin index).
- Interpret and visualize the results to identify clinically significant subgroups and explore potential implications for patient care and treatment strategies.

4 Project Evaluation

Your grade on the project will be based on the scope, depth, notebook organization, clarity of your analysis, the quality of your write-up, and the performance of the best model.

5 Presentation

- Your presentation should only describe the ”best” classification/prediction model and the ”best” clustering algorithm. Summarize the steps followed to acquire the best performance for each.
- Make an unlisted YouTube presentation and submit the unlisted video link by noon, Dec. 2nd.
- The video should be about 5-8 minutes. For a group project, both members should present.
- The presentation slides should be clearly visible and the presenter(s) are preferred to be visible, if possible.
- You may use the Bb Zoom Meeting, the Bb Collaborator Ultra, or other platform of your choice.
- Content: describe the process of building and testing the best models/algorithms from beginning to the end.

6 References

Use the IEEE citation format: numerical citations in square brackets to refer to **all resources** and provide straightforward formatting for references. See IEEE Citation Guidelines.