Machine Learning in Finance Assignment 3

Due date: 3/26/2023 (Tentative)

March 10, 2023

1 Unsupervised Learning

- Background: The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables (also see datacard.txt).

1.1 Exploratory Data Analysis

- Obtain and import the dataset (*HW3.csv*) from Courseworks as a Pandas dataframe.
- Data cleaning: Some features have missing values and outliers. Choose the way you see as appropriate to clean the dataframe and deal with outliers.
- Plot the correlation matrix of all numeric features. What do you discover?

1.2 Principal Component Analysis

- What is Principal Component Analysis? How do we interpret the "first principal component"?
- Normalize the data. Why is normalization necessary here?
- Use sklearn's PCA to find the first three principal components. What is the percentage of variance explained by each component? Do you think the dataset is well-represented in the new 3d space?

1.3 K-means clustering

- Use the data transformed by PCA, plot the Elbow curve of inertia and find the right k (set initialization to "random").
- Fit a final model and plot the clustering result as a 3d graph. Briefly discuss your findings.
- Use the data transformed by PCA, plot the Elbow curve of inertia and find the right k (set initialization to "kmeans++").
- Fit a final model and plot the final result. Briefly discuss your findings.
- Use the data before the PCA transformation and repeat Kmeans clustering (with elbow curve and "kmeans++" initialization). Select 5 variables to plot the distributions of each variable cluster by cluster. Briefly discuss your findings.
- After the above studies, what are your thoughts on the pros and cons of performing PCA before clustering?

2 NLP

- Please complete the HW3 Jupyter Notebook file for this part

3 Project 1

- Please submit a proposal for your project 1. It can be anything related to machine learning in finance.
- The proposal should not be more than 1 page long but should at least include the ML problem type, features, labels, and models used. (Check Lecture 2)