Homework VI

STAT/CS 5810/6665 - Spring semester 2024

Please upload your solutions in a single pdf file in Canvas. Any requested plots should be sufficiently labeled for full points. Include any code requested.

Unless otherwise stated, programming assignments should use built-in functions in your chosen programming language (Python, R, or Matlab).

1. **Reinforcement learning (10 pts).** Read the review paper on reinforcement learning, which you can find on Canvas. Write a short (1-2 paragraph) summary of the paper. Are there any problems where you think reinforcement learning will become particularly useful?

2. **EDA (Exploratory Data Analysis) - 50 pts**

   A superstore is planning for the year-end sale. They want to launch a new offer - gold membership, that gives a 20% discount on all purchases. It will be valid only for existing customers and the campaign through phone calls is currently being planned for them. The management feels that the best way to reduce the cost of the campaign is to make a predictive model which will predict whether existing customers will purchase the offer. Before that, we want to do some exploratory data analysis. Here we will use the "superstore_data.csv" on Canvas. Our data has 2240 values and 22 columns.
   Here is a description of the variables:
   - Response (target): 1 if the customer accepted the offer in the last campaign, 0 otherwise
   - ID: Unique ID of each customer
   - Year_Birth: Age of the customer
   - Complain: 1 if the customer complained in the last 2 years
   - Dt_Customer: date of customer's enrollment with the company
   - Education: customer's level of education
   - Marital: customer's marital status
   - Kidhome: number of small children in customer's household
   - Teenhome: number of teenagers in customer's household
   - Income: customer's yearly household income
   - MntFishProducts: the amount spent on fish products in the last 2 years
   - MntMeatProducts: the amount spent on meat products in the last 2 years
   - MntFruits: the amount spent on fruit products in the last 2 years
   - MntSweetProducts: amount spent on sweet products in the last 2 years

- MntWines: the amount spent on wine products in the last 2 years
- MntGoldProds: the amount spent on gold products in the last 2 years
- NumDealsPurchases: number of purchases made with discount
- NumCatalogPurchases: number of purchases made using catalog (buying goods to be shipped through the mail)
- NumStorePurchases: number of purchases made directly in stores
- NumWebPurchases: number of purchases made through the company's website
- NumWebVisitsMonth: number of visits to company's website in the last month
- Recency: number of days since the last purchase

- 2.1 Load the dataset and examine if the data contain missing values. If so, impute the missing values by the the method of your choice and explain why you used that method for imputation.
- 2.2 Explore the distributions of numerical variables such as Age, Income, and the amounts spent on 3 different product categories. You can either make some histograms (density plots) or calculate some summary statistics (mean, median, standard deviation, etc.). Briefly comment on what you've observed from this analysis.
- 2.3 Explore the distribution of categorical variables such as Education and Marital Status. How many unique categories are there in each variable? What are the most common categories? To answer the question, you may choose 3 different categorical variables that you are interested in and use bar plots to visualize them.
- 2.4 Explore the relationship between the number of purchases made (e.g., NumStorePurchases, NumWebPurchases) and the amount spent on different product categories. You can use any methods that you prefer.
- 2.5 Analyze the relationship between customer demographics (e.g., Age, Income) and their purchasing behavior (e.g., Number of purchases, Amount spent on different product categories).
- **(6655)** Apply one of the following visualizations methods to the data: PHATE, RF-PHATE, t-SNE, or UMAP. Note that if you do not use RF-PHATE, you will need to exclude the categorical variables. You may need to also standardize the continuous variables (note that this is not neccessary for RF-PHATE). Create 5 visualizations colored by variables of your choice (note that you can use the categorical variables for coloring).
- 2.7 Perform K-means clustering to identify distinct groups of customers based on their purchasing behavior. You may do this by determining which variables you think are associated with customer behaviors and then apply K-means on those variables. Visualize the clusters using the visualization you created before (if in 6655) or by using scatter plots or parallel coordinate plots to understand the characteristics of each cluster.
- 2.8 Explore the trend in customer visits to the company's website over time (e.g., NumWebVisitsMonth) using time series plots. Analyze changes in customer behavior over time by comparing the distribution of variables across different enrollment dates (Dt_Customer).

3. **Maching Learning - 40 pts**

Now we are tasked with performing a comprehensive comparison of different classifiers on the same dataset containing information about customer behavior and purchases. The goal is to predict whether a customer accepted the offer in the last campaign based on demographic and behavioral variables. **You can use any existing machine learning library to complete the task**.

- **3.1** Split the dataset into training, and test sets. For reproducibilty, you may specify a seed for this.

- **3.2** Model Selection and Hyperparameter Tuning:

  (5810) Choose four of the following classifiers to apply to the data. Among those four models you've chosen, ensure that there is at least one classifier that provides variable importances (e.g., Random Forest or XGBoost).

  (6655) Apply all of the following classifiers to the data:

  k-Nearest Neighbors (k-NN)
  Linear Discriminant Analysis (LDA)
  Quadratic Discriminant Analysis (QDA)
  Logistic Regression
  Support Vector Machines (SVM)
  Random Forest
  XGBoost

  For each classifier, perform hyperparameter tuning using techniques such as grid search or random search. Use either cross-validation or a holdout approach, tune hyperparameters on the training set and validate on the validation set. **Hint:** You don't have to code the cross-validation by yourself, you can use the built-in function sklearn.model_selection.GridSearchCV, here is link of this method

- **3.3** Model Evaluation:

  Evaluate the performance of each classifier on the test set using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC. Compare the performance of different classifiers and identify the best-performing model. Plot the confusion matrix for each classifier as well. Which model is the best-performing model?

- **3.4** Compute variable importances using one of your better-performing methods. Which features contribute the most to the prediction task?

- **3.5** Reporting: In your main pdf, make sure to summarize the results of the comparison, including:

  - Performance metrics of each classifier on the test set.
  - Plots of confusion matrices of each classifier.
  - Variable importance for one of the tree-based models.
  -Some insights into the predictive power of different features and their importance in predicting customer acceptance.