The Hashemite University

Prince Al-Hussein bin Abdullah II Faculty for Information Technology

Department of Information Technology

Machine Learning (2010042321)
Second Semester 2022/2023
Project

PART 1: Data Visualization

In this part you will use the California Housing Prices dataset (housingdata.csv) from the StatLib repository. This dataset was based on data from the 1990 California census. There are 10 features: longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value, and ocean_proximity for each block group in California. Block groups are the smallest geographical unit for which the US Census publishes sample data (a block group typically has a population of 600 to 3,000 people).

Note that the description of the features is the same for all groups, but each group is assigned to a different dataset. Please, download the " housingdata.csv" that is linked to your group name (e.g. group1 → is assigned “housingdata(1).csv”).

Your Task:

- Explore and visualize the data and try to find correlations among features, outliers or any pattern in the data. (Hint: Produce a scatterplot matrix which includes all of the variables in the data set.)

PART 2: Regression

In this part you will also use the California Housing Prices dataset.

Your Tasks:

- Split your dataset into train and test datasets using the hold-out method.
- Construct regression models considering the forward selection strategy.
- For each model report the: (i) R^2 score and (ii) Mean Squared Error.
- Evaluate the resulting regression models and identify the best model.

PART 3: Classification

Data sets: The data set to be used in this part of the assignment is called “Online Shoppers Purchasing intention” dataset. The dataset consists of 12330 sessions (records), each session represents a user. Each session is represented by 17 features and a class label called “revenue” which takes two values negative and positive. The negative class samples represent sessions that did not end with shopping, while the positive class samples represent sessions ending with shopping.

The description of features is as follows:

- *"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.*

- *The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.*
- *The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.*
- *The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.*

Note that the description of the features is the same for all groups, but each group is assigned to a different dataset. Please, download the " Online Shoppers Purchasing intention -Dataset" that is linked to your group name (e.g. group1 → is assigned “Online Shoppers Purchasing intention-Dataset (1)”).

Your Tasks:

Each group member should apply one classification algorithm and report the results. Classification algorithms to be used are Decision tree classifier, Bayes classifier, Lazy classifier, SVM and Ensemble classifier.

- Classify your dataset based on the 10-fold cross validation. For each individual experiment, report the confusion matrix.
- Classify your dataset based on the hold-out method. For each individual experiment, report the confusion matrix.
- Compare the classification results that were obtained when using 10-fold cross validation and when using hold out method. What did you observe?
- Explain any preprocessing technique(s) you have used, why you have used those techniques.
- Identify the classifier (Classification model) that the group has selected - the type, it's performance, reasons for selection.

Dr. Esra’a Alshdaifat