Lewis Francisco Handy-Cardenas, Matrix Methods for Machine Learning, ECE532, Project Proposal.

Project Proposal - "Online Shopper Intention Classification"

Dataset that will be used

The dataset that was chosen to be used is called "online_shoppers_intention" and can be found under https://www.kaggle.com/roshansharma/online-shoppers-intention?select=online_shoppers_intention.csv

The data consists of 18 categories, 10 of them being numerical attributes and the rest categorical attributes. One of the categories is "Revenue", which is labeled as "TRUE" or "FALSE". This would be the class label, and the objective would be to predict if a new shopper will end up buying something ("revenue" – "TRUE"). The dataset consists of around 12,000 entries.

The 18 attributes are of different kinds. There are three attributes (Administrative, Informational, Product related) which represent the different types of pages that were visited by the customer in a single session. Each of these attributes has a corresponding attribute which represents the total time spent in each of the page categories (e.g Administrative Duration). According to the data source, some of the attributes are obtained from "Google Analytics" metrics; these are the "Exit Rate", "Bounce Rate" and "Page Value". The rest of the attributes are self-explanatory, and are as follows:

- "Special day". Indicates if the customer is viewing items online for potential purchase on a holiday day such as Christmas, Black Friday, Valentine's day, etc.
- "Month". Specifies the month of the year the customer started the session on.
- "Operating system". Specifies the browser used by the customer, for which a number is analogous to some known operating system (e.g. 1 → Windows).
- "Browser". Specifies the browser used by the customer, for which a number is analogous to some known browser (e.g. 1 → Google Chrome).
- "Traffic type". An attribute related to some internal measure for which a number is analogous to some known label.
- "Visitor type". This specifies if the customer is a returning or new visitor.
- "Weekend". This specifies if the user is shopping on a weekend (Saturday or Sunday) or during the week (Monday-Friday).

The classification problem that will be explored is to predict if a user will purchase something online when browsing during a session.

Algorithms that will be applied to the dataset

Because of the type of labels (True or False), a binary classifier would be appropriate to classify the information. The following three algorithms will be applied to the dataset: Linear regression: Least squares, Page-rank algorithm, Neural Networks with multiple layers. Cross validation will be used to determine the effectiveness of each of the algorithms. Initially, hold-out data will set aside, and the remaining data will be used as training data. The starting iteration for this would be to divide the data set in 10,000 entries to be used as the training data, and the remaining 2,330 to be used for evaluation. This process will be repeated a couple of times to determine the optimum efficiency of each of the algorithms.

At first glance, the raw data shows some sparsity within some of the attributes that may cause some issues. One alternative to trying to tackle this is to use an unsupervised algorithm on the data for pre-processing before applying the supervised algorithms. Principal Component Analysis (PCA) would be used to reduce dimensionality and get rid of some of the attributes that may cause issues.

In order to benchmark the performance of each algorithm, the maximum total effectiveness of each algorithm obtained after the cross-validation process will be compared one another. If the three proposed algorithms render very poor performance and accuracy, a different algorithm from the ones reviewed in class (e.g. K-means clustering or Kernel regression) will be implemented in place to create a better classifier.

Project Github

A link to the project files can be found under the following path:

https://github.com/handycardena/ECE532_Final_Project_Handy

https://github.com/handycardena

Timeline for the progress of the project

Important dates:

October 22 → Delivery of Proposal

November 17th → First update

December 1st → Second update

December 12th → Delivery

The following Gantt diagram breaks down the project goals for the project development timeline.