Lewis Francisco Handy-Cardenas, Matrix Methods for Machine Learning, ECE532, Project Update 1.

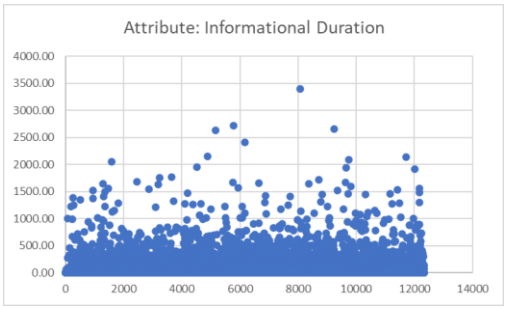
Project Update 1 - "Online Shopper Intention Classification"

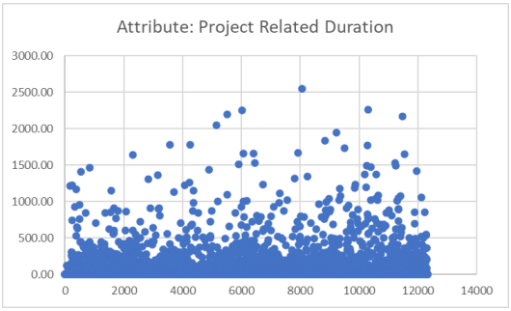
Data preprocessing

As mentioned previously, 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, and in an initial approach to start the machine learning process, the training sets were initially designed to include the first 1,000 entries. This size will later be changed and optimized according to evaluation.

For now, in order to ease the classifier implementation, only the 10 numerical attributes of the dataset are being used to train the classifiers. The complexity introduced by having multiple datatypes introduced the necessity to create labels for the features that are given as string variables; for example, the days of the month. Additional research and documentation is being conducted in means to implement the remaining categories which are given as "string" type variables. For now, the "revenue" category was binarized with respect to its "TRUE" or "FALSE" value; a value of 1 represents true, and a value of 0 represents "FALSE". This "FALSE" value will most likely be changed to -1.

Some of the data attributes contain very sparse data which complicates the calculation for the different type of machine learning techniques. The following plots show the range some of these values can be found for the attributes "Informational Duration" and "Project Related Duration".





The sparsity given in some of the values of these attributes could be conflicting with some of the machine learning techniques that are being implemented, it could well be that these attributes are not contributing a lot into the weight vector, diminishing their importance to be included in the predicting model to be developed.

Linear regression: Least squares classifier

The implementation of the least squares classifier utilizing the training data set was implemented. It has been found that if the training set is to big, the computation for the weight vector found yields NaN values. There is a slight suspicion that this could be given because of:

- Improper formulation of the label vector → change "FALSE" to -1 and not to 0.
- Some of the attributes with big variations in the data (i.e. Informational Duration and Project Related Duration) are complicating the calculation.

For now, the following have been found to be the optimal weight vector values:

```
# Classifier 1
#w = (X^T X)^(-1)X^T y
X = x_train
y = y_train
w = np.linalg.inv(X.transpose()@X)@X.transpose()@y
#A = np.linalg.inv(X@X.T)

print(np.round(w,2))
```

```
[ 0.01  0.    0.02 -0.   -0.    0.    0.04 -0.09  0.01  0.01 -0.01  0.
 -0.    0.  ]
```

The weight vector still needs to be validated by running cross-validation techniques on the available dataset, in order to determine if the current weight vector is the best one for the model.

Timeline evaluation

Because of some the complications which came up during the pre-processing step, the project is running a little bit delayed with respect to what was projected with the Gant diagram. Once all of the attributes are properly implemented, it should be relatively easy to reach the project goals stablished for the second update date.

There are still two algorithms which need to be implemented: Page-rank algorithm and Neural Networks with multiple layers. At this stage of the project there are some indicators that show that it could be that maybe one of these is not appropriate; once evaluation starts with these algorithms it will be determined if a new algorithmic approach must be included outside of these because of the structure of the dataset.

Project Github

A link to the project files (including updates) 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.