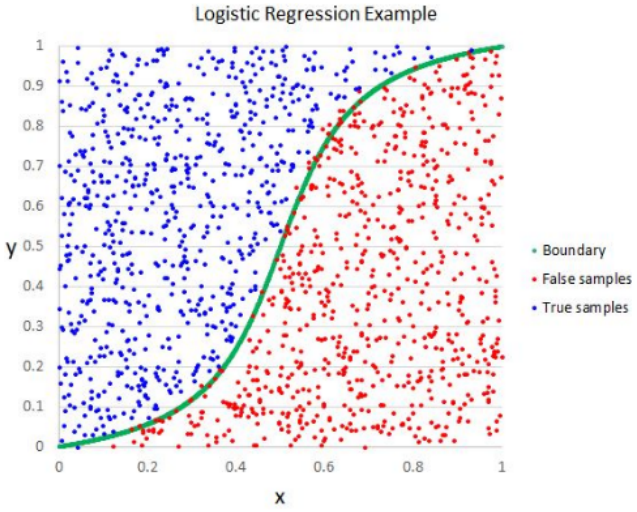
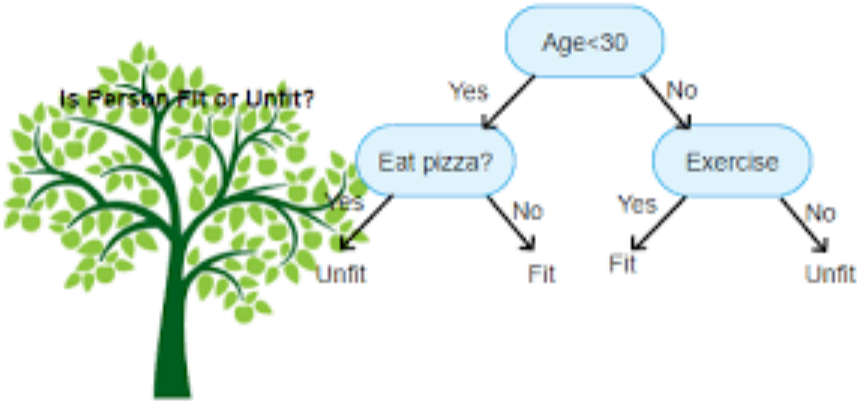
LOGISTIC REGRESSION AND DECISION TREE





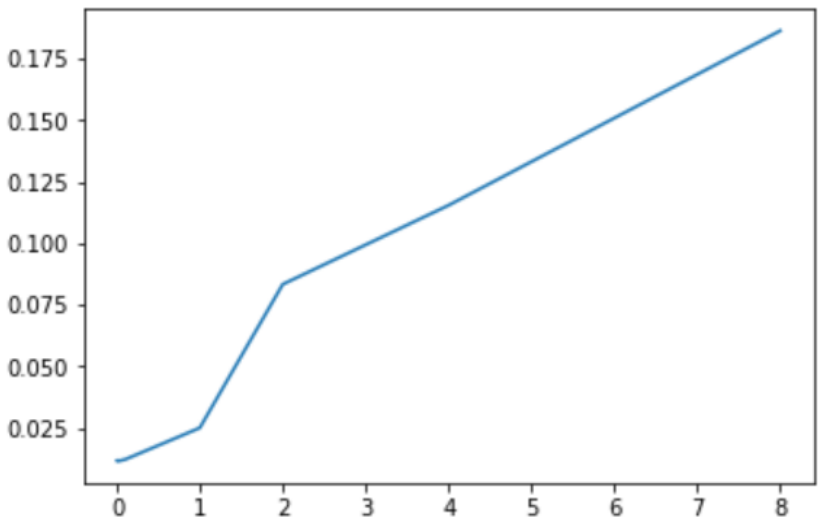
BRIJESH GOHARIA

Lab 8a:

We ran the SVM model on the spam emails dataset and wanted to find answers for the following questions:

1. The corresponding words for the top 20 positive weight from the output of the SVM model weight. Following are the words:
   '35', 'days', 'want', 'have', 'york', 'two', 'as', 'm', 'aware', 'visit', 'buy', 'advisement', 'nor', 'learn', 'economies', 'email', 'offers', 'clarify', 'think', 'presentation', 'details'.
2. The corresponding words for the top 20 negative weight from the output of the SVM model weight. Following are the words:
   'o', 'file', '200', 'deal', 'mail', 'investment', 'approval', 'early', 'which', 'ordered', 'what', 'below', 'tammie', 'they', 'predicated', 'house', 'no', 'we', 'experts', 'smoothing', 'bargain'.

3. We plotted various values of 'regParam' of training and testing error as function. We observed that at larger values of regParam, the test error increases significantly and at lower values of regParam, the testing error have less difference in values. Following is the graph:



Lab 8b:

ISLR" R" library Data set:

We ran the ISLR credit card data set to find the students who would default. Following are the process for data pre-processing and the key observations.

Hyper-parameter search:

- With the default parameter, we ran the model but the training error was "0" along with all the prediction being equal to 0
- We than changed the parameters such as **impurity='entropy'**, **maxDepth=5**, **maxBins=20 (Decision Tree)** and **iterations=10 (Logistic Regression)**

The effect of normalization:

- Taking Log of Balance and income did not affect the accuracy to any good extent

Interpretation of the results:

Logistic Regression:

```
Training Error = 0.0350819129571
Test Error = 0.0535759413697
Accuracy = 94.642405863
```

Decision Tree:

```
Test Error = 0.0338640384129
Accuracy = 96.6135961587
```

Spam-Email Data set:

Similarly, we ran the Spam-Email data set on Logistics Regression and Decision tree to classify spam vs Ham emails. Following are the process for data pre-processing and the key observations.

The effect of normalization:

- Without normalization the accuracy of the models reduced drastically by 15-20 percent

Interpretation of the results:

Logistic RegressionwithSGD: `Accuracy = 91.76628810520025`
Logistic RegressionwithLBFGS: `Accuracy = 98.75971309025702`

Categorical Features in Decision Trees (DT):
In order to work with the categorical features. We used the titanic dataset. The task was to correctly classify the people who survived based on Age and various other factors.
Data Pre-Processing:

- Imputed all the NA in the data sets
- Converted all the categorical variables into numerical as follows:

- SEX: {'female': 0, 'male': 1}
- Titles: {"Mr": 1, "Master": 2, "Mrs": 3, "Miss": 4, "Rare": 5}
- Embarked: {'S': 0, 'C': 1, 'Q': 2}
- Fare: ['Fare'] <= 7.91, 'Fare'] = 0 ; Fare > 7.91 & Fare <= 14.454➔ 'Fare' = 1
  ['Fare'] > 14.454) & ['Fare'] <= 31➔ 'Fare' = 2
  ['Fare'] > 31 ➔ 'Fare' = 3
- Age: Age<=16 ➔0 ; 'Age' > 16) & 'Age' <= 32 ➔'Age' = 1
  ['Age'] > 32) & ['Age'] <= 48) ➔'Age'] = 2; ['Age'] > 48) & (t['Age'] <= 64) ➔ 'Age'] = 3
  ['Age'] > 64 ➔'Age'=4

- Accuracy:
  Created the Labeled output and changed the hyper parameter to get the accuracy of
  84.7363 percent

Branching is as Follows:

"Title <= 1.5" corresponds to "Mr." title