Data Mining Assignment

DECISION TREE AND DESCISION BOUNDRIES

Prashant Pathak
MT19051
DMG
29-Sep,2019

Contents

Decision Tree

A decision tree is a powerful model to perform classification.

It has following main features:

- Internal nodes are test decision on the attributes of data.
- Leaf nodes are the class label.
- It selects the best attribute at each node to perform the split on that attributes.
- For selecting the best node, it uses “Gini index” or “Gain ratio” type of measures.

About the data

The data which is provided contains around 275 columns out of which last column is output label. Data contains lots of missing values, so following operation were applied on the data to make it in correct format.

1. If a column contains more than 50% of its values as ‘Null’ or None than that column is dropped because it will not give us any useful insight about the output column
   After doing so only 139 columns were remaining.
2. Now after analyzing I found column with name 'Num Pin Dot Pattern Views' which have all the values as “ALL” so I also dropped that column
3. 'Session Browser Family' and 'Session Browser Family Top 3 ' have similar data so I dropped the 'Session Browser Family' column.
4. After this in all the column which have “Null” or “None” values I changes them into “Fill None” to give them a common category.

After dropping all these columns 137 column remained.

1. Next we have to do label ending on ordinal data and one hot encoding on nominal data.
2. So, we found nominal attributes as following:
   'Session First Request Day of Week', 'Session Last Request Day Of Week', 'Assortment Level 2 Path Last', 'Content Level 2 Path Last', 'Session Browser Family Top 3', 'Session First Template Top 5', 'Session Last Template Top 5','Session First Referrer Top 5'

All these columns are one hot encoded and rest are label encoded. After doing all these steps all our data is in numeric and does have a mission value. So, our data is ready for applying Models.

Decision Tree Graph

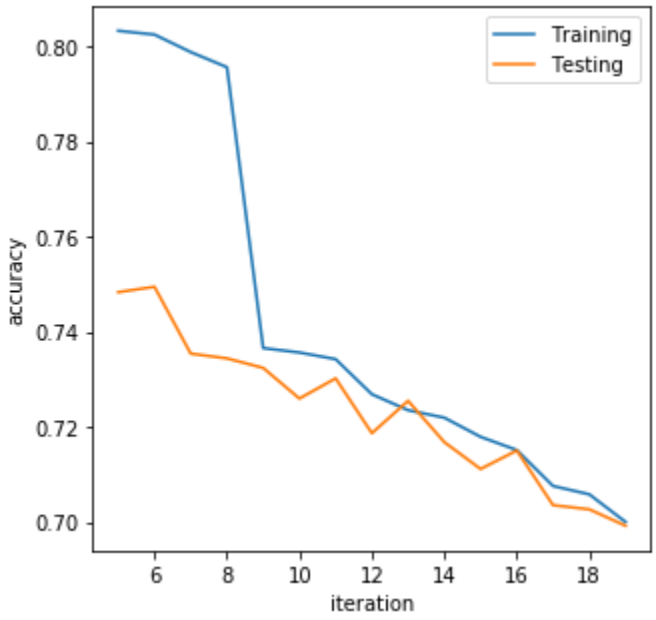

Figure 1 Accuracy vs Depth decision tree graph

In this graph we can see that our accuracy is decreasing as we increase the depth.

In testing data, the best accuracy is coming is when the dept is 6. So, we will use this depth in our next question.

Decision Boundaries.

To visualize our data set we need to convert it into two dimensions. For this purpose, we have chosen PCA, which is basically used to do dimensionality reduction.

But this data set as all the values are categorical so it might be possible that our data might not be correct.

But still we are proceeding with PCA.

The PCA diagram is shown in the next graph.

Now we will try to visualize the decision boundary of different Model. Different mode is

1) Decision Tree
2) KNN
3) Naïve Bayes

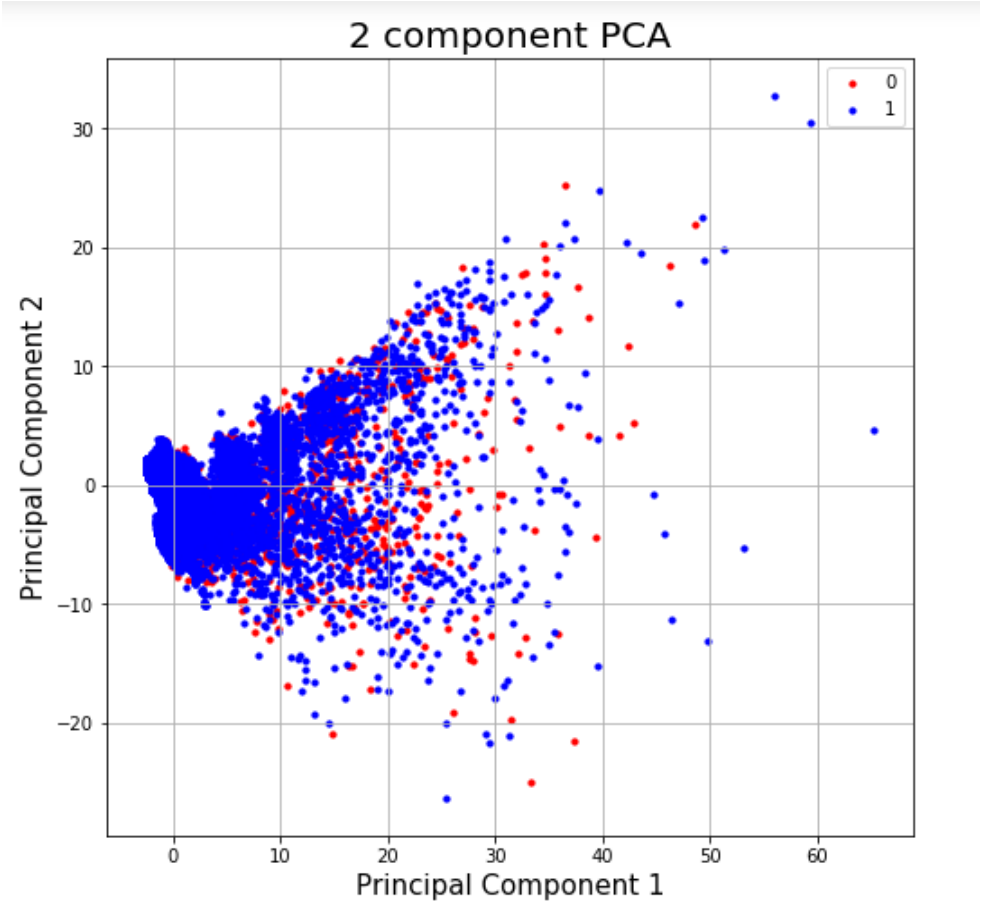

Figure 2 Visualize the dataset after PCA

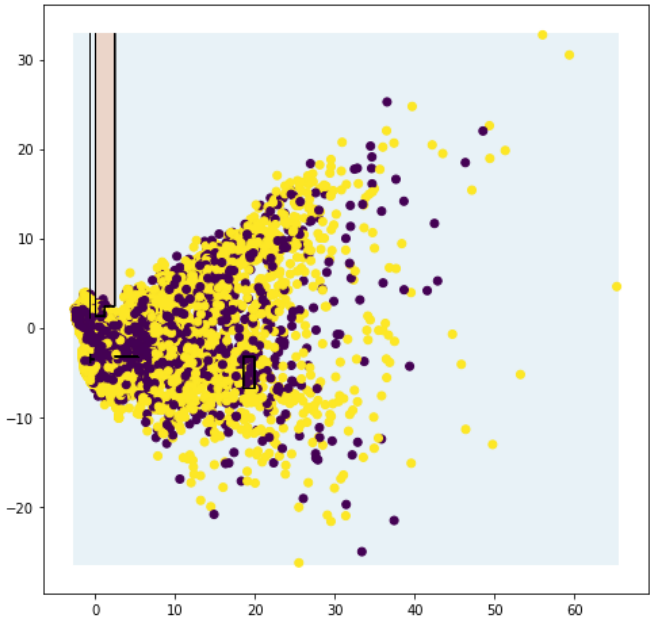

Figure 3 Decision Tree Decision Boundary

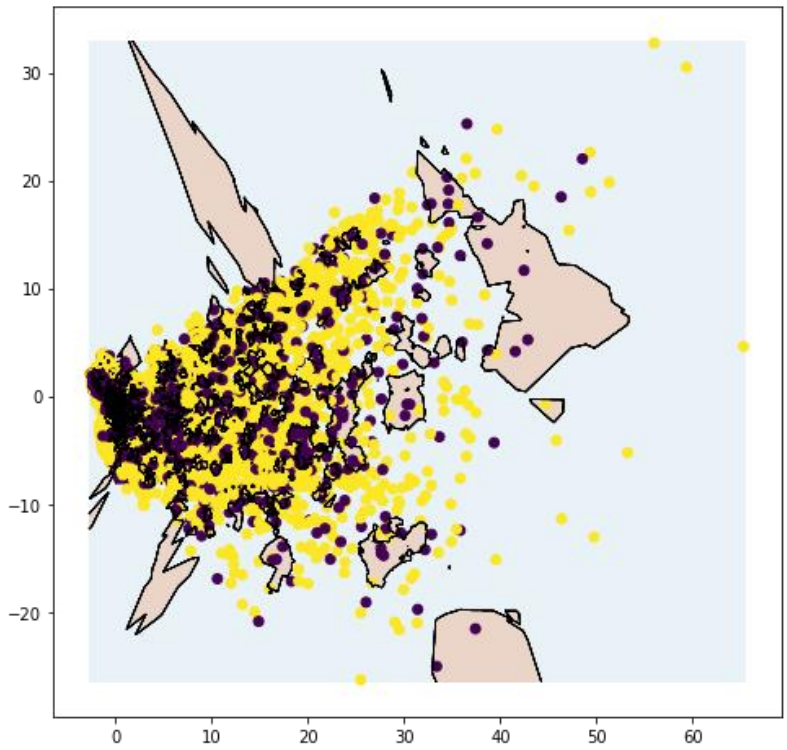

Figure 4 KNN decision Boundary

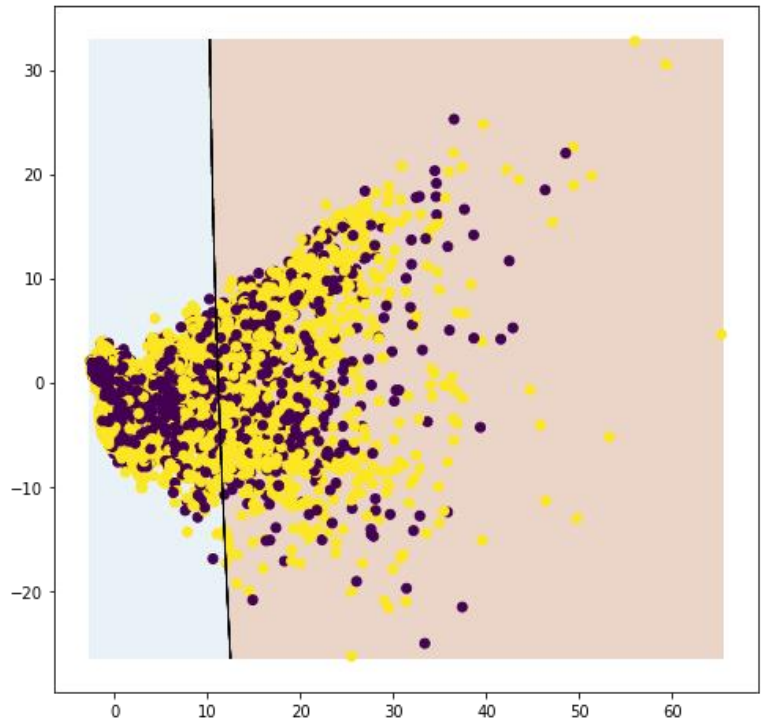

Figure 5 Naive Bayes Boundary

We can see that I our decision tree are not coming correct because our output variable is highly unbalanced. So, we will try making decision boundary with less point.

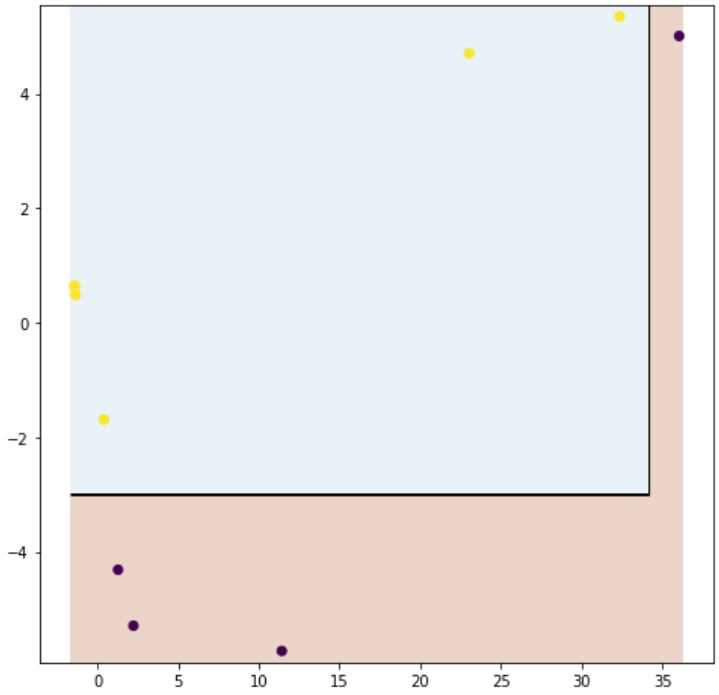

Figure 6 Decision Tree Boundary with less point

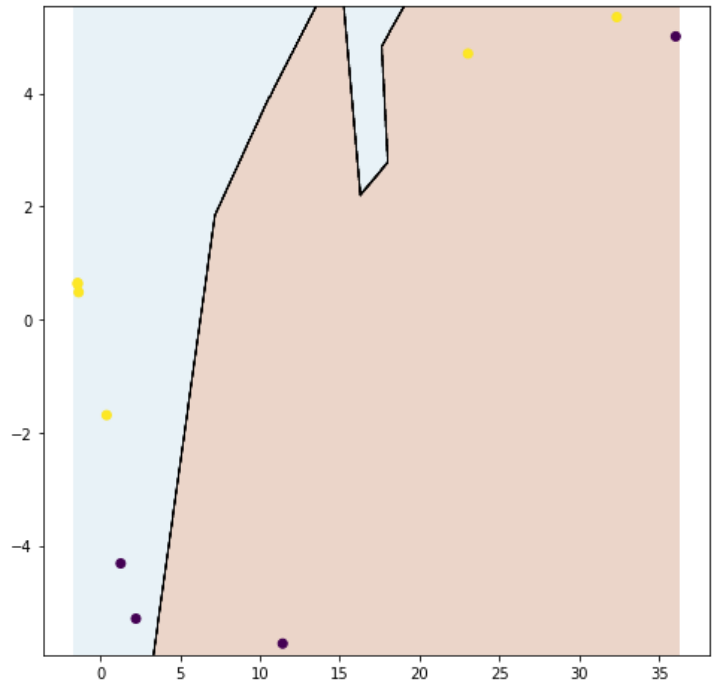

Figure 7 KNN decision Boundary with less point

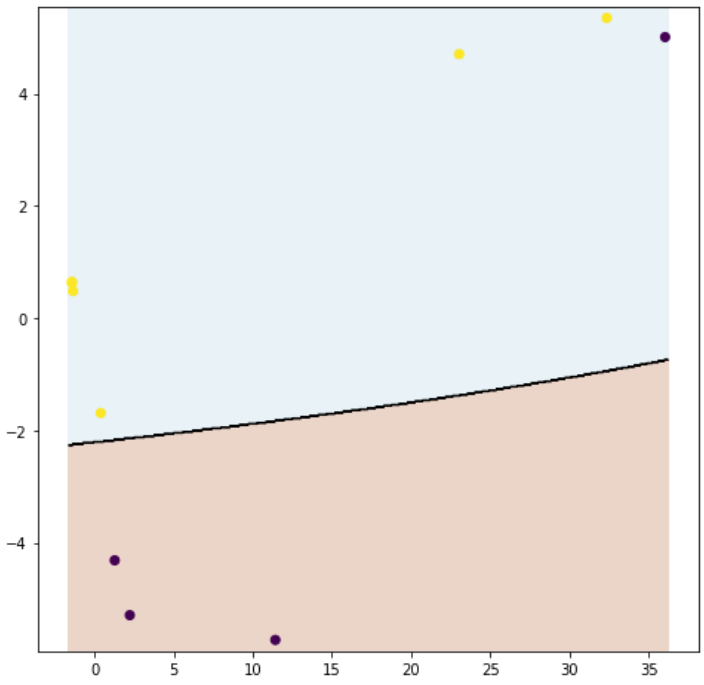

Figure 8 Naive Bayes Boundary

From the above figure we conclude that

a) Decision tree have straight line boundary and they cannot separate complex mixture of data.
b) KNN have irregular shape boundary and can separate complex data very well but it takes lots of time.
c) Naïve Bayes decision boundary is sort of slant line.

References

1) https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.score
2) https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
3) http://www.cse.chalmers.se/~richajo/dit866/lectures/l3/Plotting%20decision%20boundaries.html
4) https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
5) https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes
6) https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/