Binary Classification

Objective

Simulate a binary classi_cation dataset with a single feature via a mixture of normal distributions using R (Generate two data frames with the random number and a class label, and combine them together). The normal distribution parameters (using the function rnorm) are(5,2) and (-5,2) for the pair of samples - determine an appropriate number of samples. Induce a binary decision tree (using rpart), and obtain the threshold value for the feature in the _rst split. How does this value compare to the empirical distribution of the feature? How many nodes does this tree have? What is the entropy and Gini at each? Repeat with normal distributions of (1,2) and (-1,2). How many nodes does this tree have? Why? Prune this tree (using rpart.prune) with a complexity parameter of 0.1. Describe the resulting tree.

Solution Implemented

Two normal distributions were used to simulate binary classification data:
samp1 = rnorm(n=100, mean=5, sd=2) this normal distribution was assigned label 1
samp2 = rnorm(n=100, mean=-5, sd=2) this normal distribution was assigned label 0

Binary decision tree was created using rpart:
tree1 = rpart(Y~X, data = data_df, method = "class")
Following tree was obtained

```
> tree1
n= 200

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 200 100 1 (0.5000000 0.5000000)
  2) X>=0.07269424 100   0 1 (1.0000000 0.0000000) *
  3) X< 0.07269424 100   0 0 (0.0000000 1.0000000) *
```

With above results, the first split was made at X < 0.07269424.

Empirical distribution of the feature states its
median at 0.07269,
1st Quartile at -4.84610,
3nd Quartile at 4.97600
The number of nodes in tree are 3. Moreover, the split has been made almost at median of distribution

Root node:

Gini Index: 0.5
Entropy: 0.6931

Right node (X < 0.4846688):

Gini Index: 0
Entropy: 0

Left node (X>=0. 4846688):
Gini Index: 0
Entropy: 0

Again, two normal distributions were used:
samp1 = rnorm(n=100, mean=1, sd=2) this normal distribution was assigned label 1
samp2 = rnorm(n=100, mean=-1, sd=2) this normal distribution was assigned label 0

Binary classification tree on the combination of the above two samples resulted in 15 nodes in total as shown below :

```
> tree2
n= 200

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 200 100 1 (0.50000000 0.50000000)
    2) X>=1.092536 64  11 1 (0.82812500 0.17187500) *
    3) X< 1.092536 136  47 0 (0.34558824 0.65441176)
      6) X>=-2.24734 114  46 0 (0.40350877 0.59649123)
       12) X>=-1.134767 77  35 0 (0.45454545 0.54545455)
         24) X< 0.9906557 70  34 0 (0.48571429 0.51428571)
           48) X>=0.7075104 8   1 1 (0.87500000 0.12500000) *
           49) X< 0.7075104 62  27 0 (0.43548387 0.56451613)
             98) X< -0.4339205 30  13 1 (0.56666667 0.43333333)
              196) X>=-0.5727713 7   0 1 (1.00000000 0.00000000) *
              197) X< -0.5727713 23  10 0 (0.43478261 0.56521739) *
             99) X>=-0.4339205 32  10 0 (0.31250000 0.68750000) *
         25) X>=0.9906557 7   1 0 (0.14285714 0.85714286) *
       13) X< -1.134767 37  11 0 (0.29729730 0.70270270) *
      7) X< -2.24734 22   1 0 (0.04545455 0.95454545) *
```

Greater number of nodes were received here because the two samples were partially overlapping in the sample space. The tree was grown to improve the classification error rate at each child node till improvement increases by not more than default complexity parameter(0.01).

After pruning the tree with a complexity parameter of 0.1, we got the following tree with 3 nodes:

```
> tree3
n= 200

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 200 100 1 (0.5000000 0.5000000)
  2) X>=1.092536 64  11 1 (0.8281250 0.1718750) *
  3) X< 1.092536 136  47 0 (0.3455882 0.6544118) *
>
```

Split was done at X < 1.092536

Right child node at X<1.092536:
Predicted class = 0, observations with class 1 = 47, observations with class 0 = 89
Probability of class 1 = 0.346
Probability of class 0 = 0.654

Left child node at X>=1.092536::

Predicted class = 1, observations with class 1 = 53, observations with class 0 = 11
Probability of class 1 = 0.828
Probability of class 0 = 0.172