HW2: Supervised Learning

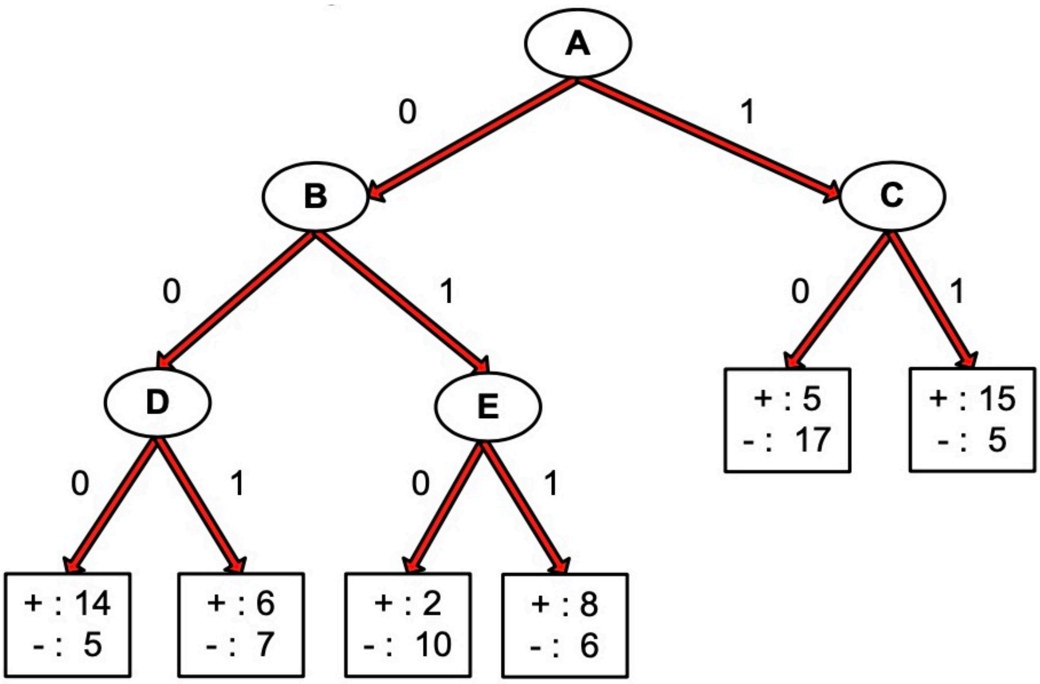
Name: Prudvi Kamtam
UCF ID: 5498416

Task 1:
https://github.com/greysou1/ML-assignments/blob/main/HW2/task1.ipynb

Task 2:
https://github.com/greysou1/ML-assignments/blob/main/HW2/task2.ipynb

A) Decision tree training error
The training error rate for the given tree is 0.29,
- The training error rate of a decision tree is given by the number of wrongly classified values divided by total number of values for all the leaf nodes

$$\text{mean}\left(\frac{\text{Number of Wrongly classified values}}{\text{Total number of values used in training}}\right)$$



- The class which has the least number of values classified is the wrongly classified class and the class with the majority of the values classified is the output classification of that leaf node

Leaf Node 1:
wrongly_classified = '-', total items = 19, number of wrongly classified = 5,

Leaf Node 2:
wrongly_classified = '+', total items = 13, number of wrongly classified = 6,

Leaf Node 3:
wrongly_classified = '+', total items = 12, number of wrongly classified = 2,

Leaf Node 4:
wrongly_classified = '-', total items = 14, number of wrongly classified = 6,

Leaf Node 5:
wrongly_classified = '+', total items = 22, number of wrongly classified = 5,

Leaf Node 6:
wrongly_classified = '-', total items = 20, number of wrongly classified = 5,

Total misclassified = 5 + 6 + 2 + 6 + 5 + 5 = 29
Total data points used for training = 19 + 13 + 12 + 14 + 22 + 20 = 100

Error rate of the decision tree = 29/100 = 0.29

B) Test instance: T= {A=0, B=1, C=1, D=1, E=0}

The tree traversal path for the given test instance is
A–> B–> E–> '–'

Thereby classifying as negative (-ve)

Step 1) A = 0
The left sub-tree is selected

Step 2) B = 1
The right sub-tree is selected

Step 3) C = 1
Since, the subtree {'E'} does not have the node 'C', we will skip this

Step 4) D = 1
Since, the subtree {'E'} does not have the node 'D', we will skip this

Step 5) E = 0
The left sub-tree is selected.

Step 6) Leaf-node

Since this is a leaf-node, the class with the majority of values classified is the output of this leaf node. {‘+’: 2, ‘-‘: 10}.

Since, ‘-‘ (-ve) has the most number of values, it is selected as the output classification.

Task 3:
Splitting:

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Q1: Overall Gini before splitting = 0.48

Q2: Gain after splitting on A = 0.136

Q3: Gain after splitting on B = 0.1638

Q4: The decision tree would choose B as the root node as it has the higher gain.

The calculation on next page.

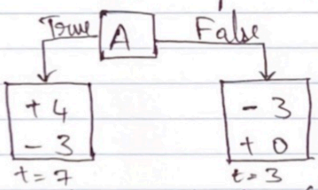
Q1: Overall gini before splitting

+	4
−	6

$$Gini = 1 - \sum_j P(j/t)^2$$

$$= 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.48$$

Q2: Gain in gini after splitting on A



$$Gini\,1\,(True) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.491$$

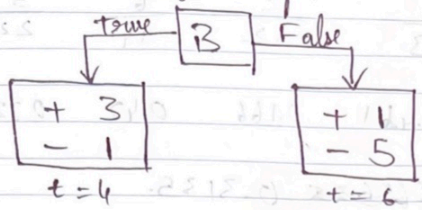
$$Gini\,2\,(False) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right) = 0$$

Gain = ~~0.48~~ Impurity before splitting − Impurity after splitting

$$= 0.48 - \left[\frac{7}{10}(0.491) + \frac{3}{10}(0)\right]$$

$$= 0.48 - 0.3437 = 0.136$$

Q3: Gain in gini after splitting on B



$$Gini\,1\,(true) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$= 1 - 0.5625 - 0.0625$$

$$= 0.375$$

$$Gini\,2\,(False) = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2$$

$$= 1 - 0.028 - 0.695$$

$$= 0.277$$

$$Gain = 0.48 - \left[\frac{4}{10}(0.375) + \frac{6}{10}(0.277)\right]$$

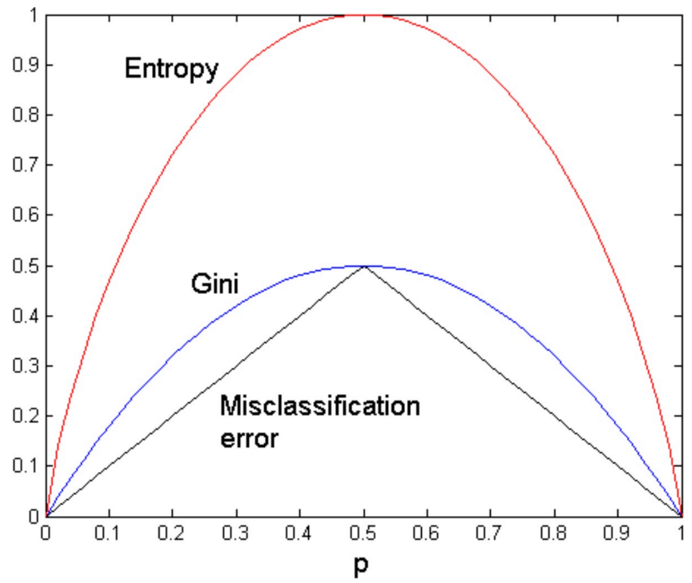
$$= 0.1638$$

Task 4:

Q1: Decision trees are not linear classifiers. This is because

- It can be used for both classification and regression of non-linear data.
- It cannot be expressed as an equation as there is no relation between dependent and Independent variable:
- It can classify data that cannot be classified using a single decision boundary unlike linear models models.

Q2: Misclassification error vs Gini index

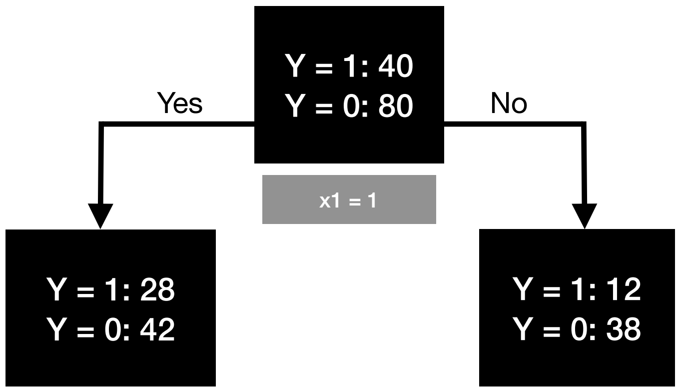


Gini index is better than Misclassification error because:

- Misclassification error sometimes isn't able to capture information gain even though the Gini index reports information gain

- This causes the algorithm to miss some of the features to be considered for the root node.

- To demonstrate this lets look at an example below



Misclassification Error:

$$I_E(t) = 1 - max\{p_i\}$$

Misclassification Error at parent: 40/120
Misclassification Error at child 1: 28/70
Misclassification Error at child 2: 12/50

=> Information gain after splitting using Misclassification Error: 0

Gini Index:

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

Gini at parent: 0.444
Gini at child 1: 0.48
Gini at child 2: 0.3654

=> Information gain after splitting using Gini Index: 0.012

We can see that even though the information gain given using Gini is 0.012, information gain given using Misclassification Error is 0. Therefore, Misclassification error is not a better spilling criteria for decision trees than Gini index.

Task 5:
Weakness of Bagging:

1. It requires many decision trees to eliminate bias.
2. It is computationally expensive.

Difference between Bagging and Random Forest:

1. In bagging, all of the features are used for training multiple Decision Tree models after sampling with replacement
2. In Random Forest, always m < M features are selected for training each model.
3. Here, 'm' is a randomly selected subset of features from 'M' total features.

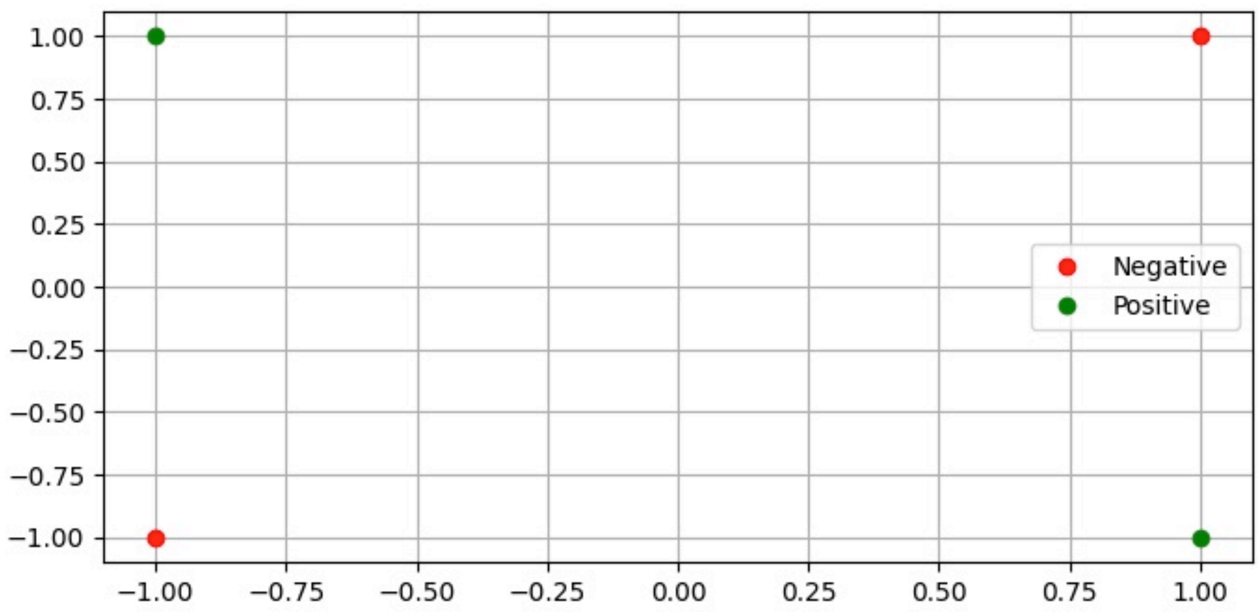
Overcoming the weakness of bagging:

1. Since, only a subset of features are used in each model, the model creation process is faster.
2. This reduces the dimensionality of multiple datasets on which a different models is trained
3. This results in a computationally faster training process as each model's training is done using only a subset of features.

Task 6:
https://github.com/greysou1/ML-assignments/blob/main/HW2/task6.ipynb

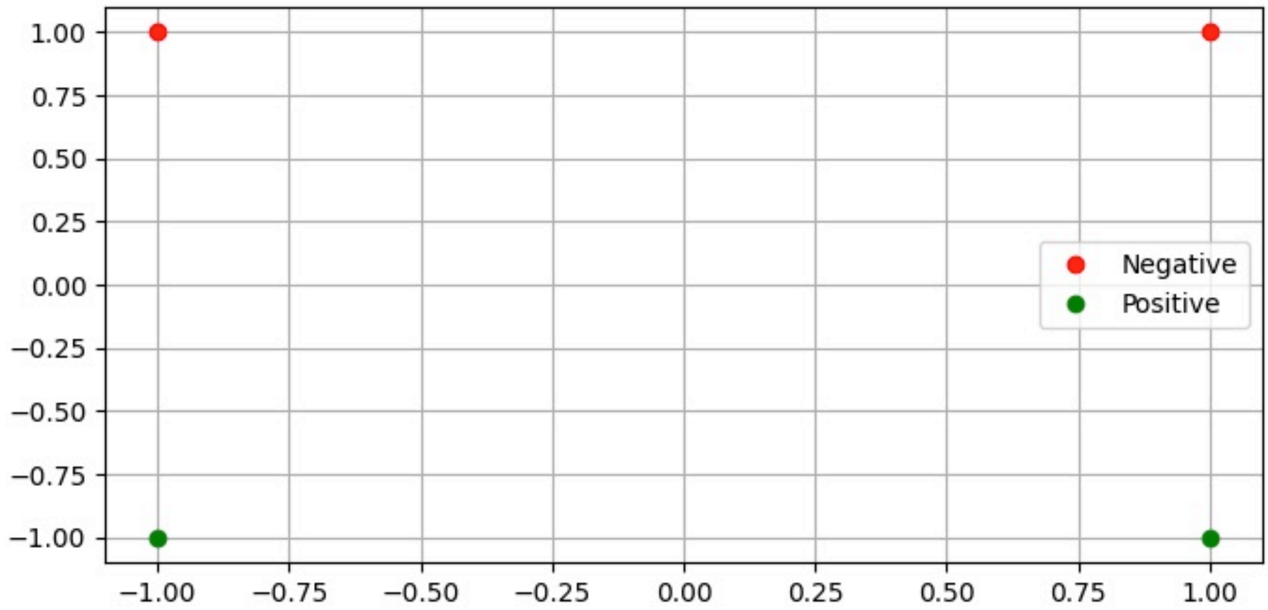
The following are the given points mapped in space:

```
[−1, −1] (negative)
[−1, +1] (positive)
[+1, −1] (positive)
[+1, +1] (negative)
```

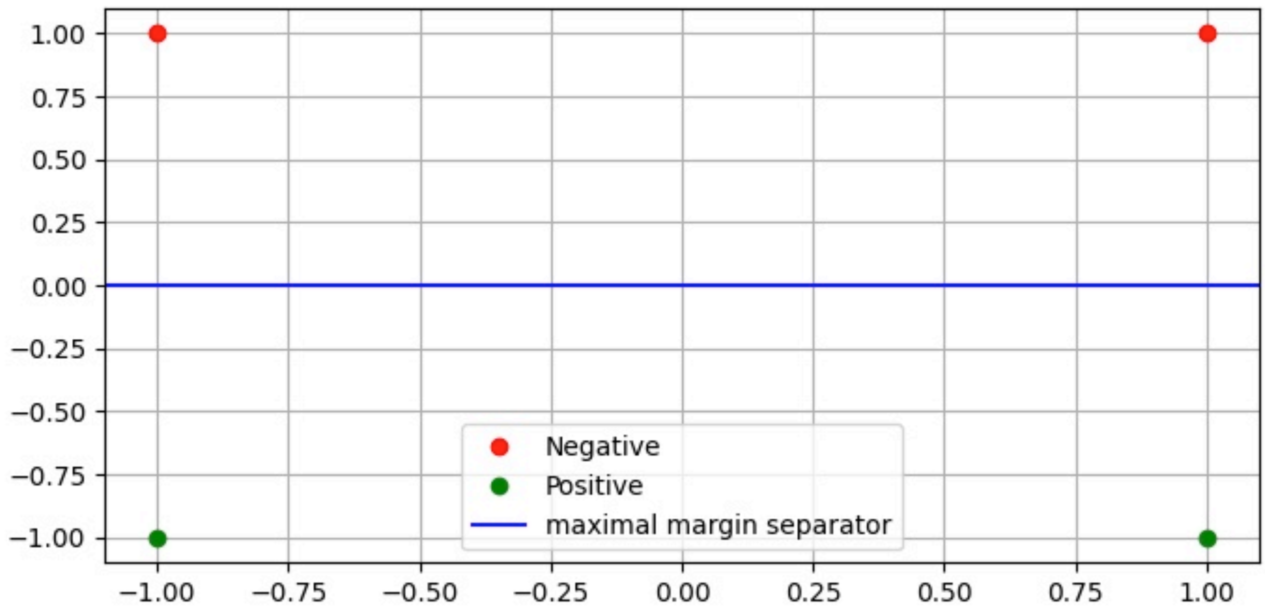


As we can see, these points cannot be separated by a single linear line, therefore the given inputs [x1, x2] mapped into [x1, x1x2] are as follows:

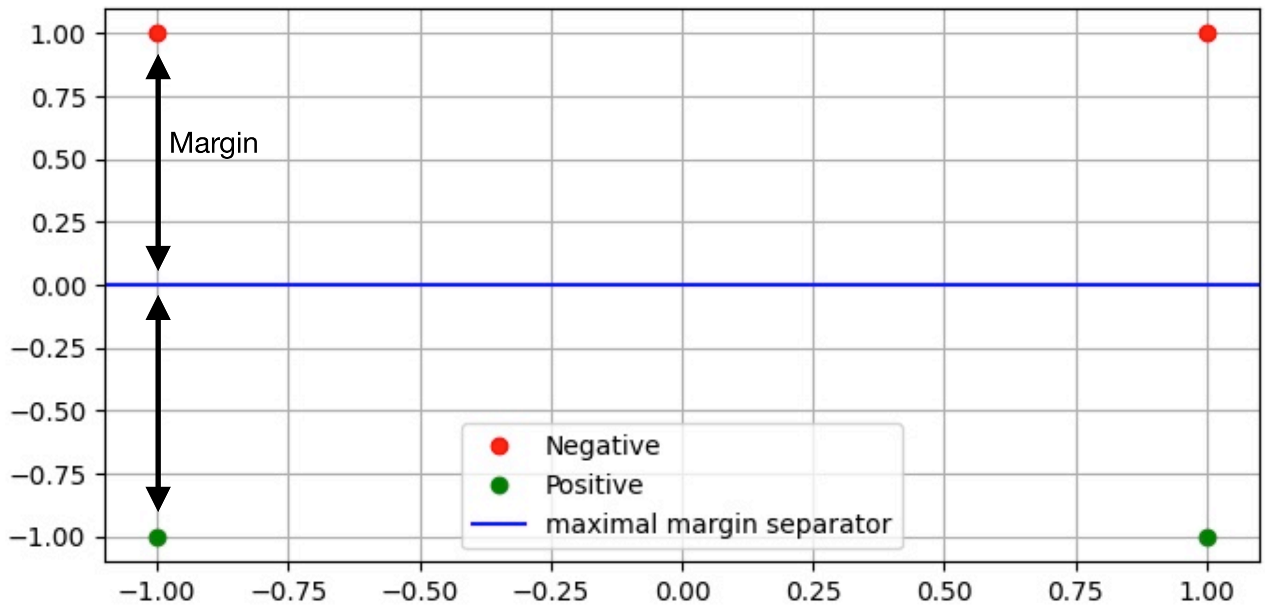
```
[-1, -1] -> [-1,  1]
[-1,  1] -> [-1, -1]
[ 1, -1] -> [ 1, -1]
[ 1,  1] -> [ 1,  1]
```



Now, these points can be clearly separated by a decision boundary (maximal margin separator) shown below:



The margin is the distance from the decision boundary to either of the support vectors.



The distance from the support vectors ex: (-1, 1) to the margin separator is the margin.

The margin is 1.0 units.

Task 7:

The circle equation:

$(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$

expands out to the following:

$\Rightarrow x_1^2 + a^2 - 2ax_1 + x_2^2 + b^2 - 2bx_2 - r^2 = 0$

This can be re-arranged as following:

$\Rightarrow x_1^2 + x_2^2 - 2ax_1 - 2bx_2 + (a^2 + b^2 - r^2) = 0$

This equation corresponds to the following:

W.X + c = 0

Where,

W = (2a, 2b, 1, 1) and

C = $(a^2 + b^2 - r^2)$

∴ The circle equation is linearly separable in the feature space (x_1, x_2, x_1^2, x_2^2).

Task 8:

The ellipse equation:
c(x1 – a)2 + d(x2 – b)2 – 1 = 0

expands out to the following:

$\Rightarrow c(x_1^2 + a^2 - 2x_1a) + d(x_2^2 + b^2 - 2x_2b) - 1 = 0$

$\Rightarrow cx_1^2 + ca^2 - 2cx_1a + dx_2^2 + db^2 - 2dx_2b - 1 = 0$

This can be re-arranges as following:

$\Rightarrow cx_1^2 + dx_2^2 - 2acx_1 - 2bdx_2 + a^2c + b^2d - 1 = 0$

This equation corresponds to the following:

W.X + c = 0

Where,
W = (2ac, 2bd, c, d, 0) and
C = $(a^2 + b^2 - 1)$

∴ The ellipse equation is linearly separable in the given feature space.