CSE 572: Data Mining

HW-2

Name: Dev A Patel

ASU ID: 1229417087

Task 1 (20 points) For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the individuals from the test dataset had survived or not. Please:

1) **Preprocess your Titanic training data;**
2) **(5 points) Learn and fine-tune a decision tree model with the Titanic training data, plot your decision tree;**
3) **(5 points) Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy;**
4) **(5 points) Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy;**
5) **(5 points) Which algorithm is better, Decision Tree or Random Forest? What are your observations and conclusions from the algorithm comparison and analysis?**

For code refer to this link: https://github.com/dpate243/CSE572-Data-Mining/blob/main/cse572_hw2_titanic.ipynb

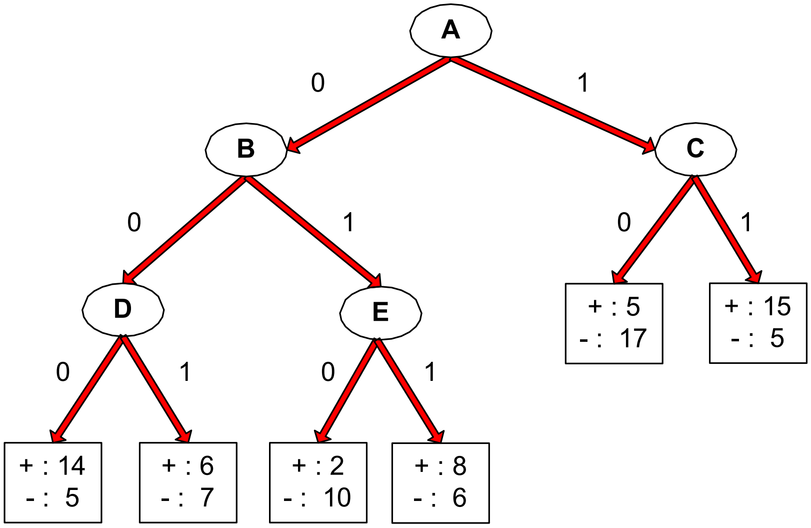
3 - Average classification accuracy after applying five-fold cross validation on Decision Tree model: **80.81%**

4 - Average classification accuracy after applying five-fold cross validation on Random Forest model: **81.59%**

5 - Random Forest performed slightly better than Decision Tree in terms of average accuracy. This is due to ensemble approach of Random Forest classifier. Decision trees are prone to overfitting when they have high depth, and hence they will not be able to predict well on unseen data. On the other hand, RF combines outputs of multiple DT mitigating problem of overfitting.

RF are more stable as compared to DT because they aggregate results from many trees. So in conclusion, Random forest should be considered over decision tree as it proves to be advantageous for our case.

Task 2 (15 points) Understanding Training Error and Testing
Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.



(a) (10 points) What is the training error rate for the tree? Explain how you get the answer?

Ans: 29%

(b) (5 points) Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

Ans: T = -

HW-2

② a) We need to find the training error rate.

for a leaf node, error is equivalent to min no. of samples, that node has.

Leaf nodes	Error
D = 0	5
D = 1	6
E = 0	2
E = 1	6
C = 0	5
C = 1	5
	29

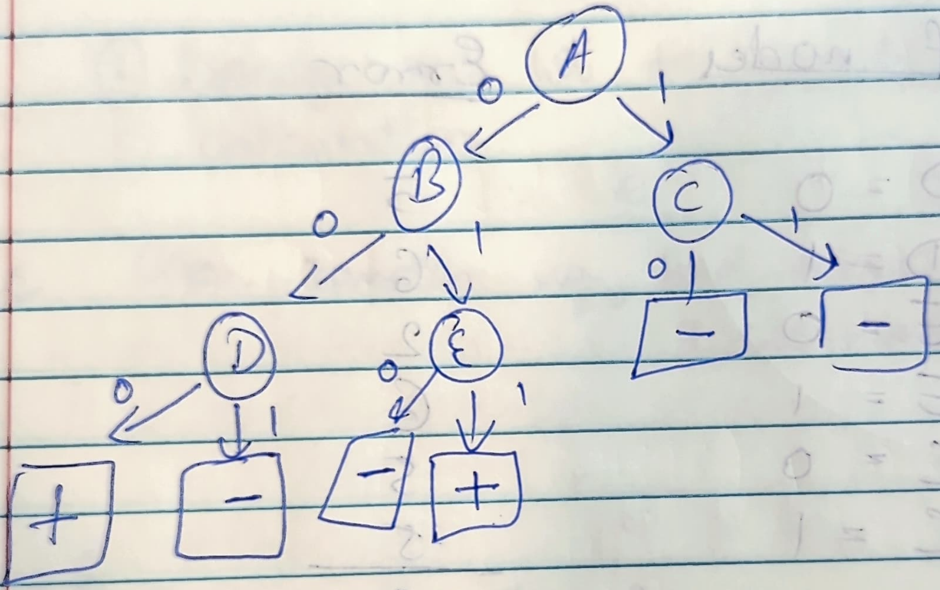
Total samples $= (14+5)+(6+7)+(2+10)+(6+8)+(5+17)+(15+5)$

$= 100$

Error rate $= \frac{29 \times 100}{100} = 29\%$

b) $T = \{ A=0, B=1, C=1, D=1, E=0 \}$

for this we need to trace the tree to a leaf node.

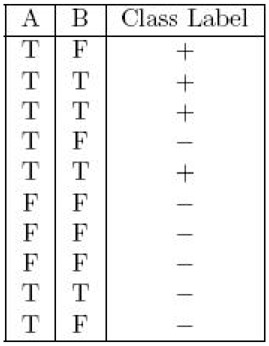


$$\boxed{A=0} \rightarrow \boxed{B=1} \rightarrow \boxed{E=0} \Rightarrow \boxed{-}$$

Values of C & D don't matter as they don't fall into the path

$$\boxed{T = -}$$

Task 3 (20 points) Understand Splitting Process
Consider the following data set for a binary class problem.



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: (5 points) What is the overall gini before splitting?

Ans: 0.48

Q2: (5 points) What is the gain in gini after splitting on A?

Ans: 0.1371

Q3: (5 points) What is the gain in gini after splitting on B:

Ans: 0.1632

Q4: (5 points) Which attribute would the decision tree choose?

Ans: B

3

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?

$$G = 1 - (0.4)^2 - (0.6)^2$$

$$= \underline{\underline{0.48}}$$

Q2 Gini gain after splitting on A. ?

gain = Impurity (P) - Impurity (A)

Impurity (P) = 0.48

$$\text{Impurity}(A) = \sum_{i=1}^{K} \frac{n_i}{n} \text{ Impurity}(i)$$

$$= \frac{7}{10}\cdot\left(G_{(A=T)}\right) + \frac{3}{10}\left(G_{(A=F)}\right)$$

$$1-\left(\frac{4}{7}\right)^2-\left(\frac{3}{7}\right)^2 \qquad\qquad 1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2$$

$$= 0.4898 \qquad\qquad = 0$$

$$\Rightarrow I(A) = (0.7 \times 0.4898) + (0 \times 0.3)$$

$$= 0.3429$$

$$\text{Gain} = 0.48 - 0.3429$$

$$= 0.1371 \longrightarrow (\text{Ans})$$

Q3 gain in gini index after split on B

$$\text{gain} = I(P) - I(B)$$
$$= 0.48 - I(B)$$

$$I(B) = \frac{4}{10} \cdot G(B=T) + \frac{6}{10} \cdot G(B=F)$$

$$= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 \qquad 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2$$

$$= 0.375 \qquad = 0.278$$

$$I(B) = 0.15 + 0.1668$$
$$= 0.3168$$

$$\text{gain} = 0.1632$$ (Ans)

Q4 Which attribute decision tree choose?

DT will choose B as it has higher gini gain than 'A'

Also, B has less gini impurity.

Task 4: (10 points) Please answer and explain.
Q1: (5 points) Are decision trees a linear classifier? Why?

No, decision trees are not linear classifiers. DT creates a model by splitting the data into subsets based on a certain condition for a particular feature. It keeps on doing it until it reaches least error or specified tree depth. Hence, it allows non-linear relation between features and target. Each leaf node in DT represents an outcome. If we take a linear classifier, it would try to find a single linear boundary to separate classes, but DTs creates multiple boundaries by splitting feature space into boxes, and hence allowing complex decision plane.

Q2: (5 points) Is Misclassification error better than Gini index as the splitting criteria for decision trees? Why?

No, it actually depends on the data over which they are to be applied. But generally gini impurity is preferred over misclassification error as ME only takes into account the final class predicted and hence it is biased towards larger class. So if the dataset is imbalanced,ME may not capture the best split as it is less sensetive to distribution of data. On other hand gini is sensetive to node impurity.

Task 5: (10 points) What are the weaknesses of bagging? What is the difference between bagging and random forests, and why such difference can overcome the weaknesses of bagging?

Weakness of bagging:

1) Each model is trained on different subset of data but uses all features, creating correlated trees.

2) High computational cost to train multiple large trees

3) More complex and hard to interpret

4) If base DT are stable with low variance, bagging will result in collection of similar trees and hence limitins improvement.

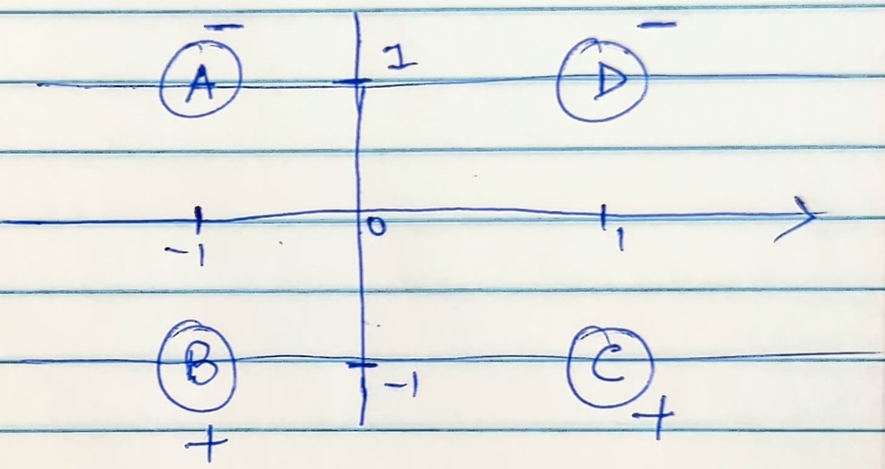
Random forest on other hand uses subset of features for each tree it creates and developing de-correlated trees. As trees are less corelated, it reduces variance effectively compared to bagging and even lead to more accurate and stable predictions. The combination of bootstrapping and feature randomization helps prevent overfitting and achieve better accuracy when compared to bagging alone.

Task 6: (20 points) Construct a support vector machine that computes the kernel function. Use four values of +1 and -1 for both inputs and outputs:
[−1, −1] (negative)
[−1, +1] (positive)
[+1, −1] (positive)
[+1, +1] (negative).
Map the input [x1, x2] into a space consisting of x1 and x1x2. Draw the four input points in this space, and the maximal margin separator. What is the margin? 【To be consistent with our lecture notes, margin is defined as the distance from the middle way/hyperplane to either support vectors. 】

Ans: Length of margin = 1 unit

6.

	Points	Map to (X_1, X_1X_2)	Sign
A	$(-1, -1)$	$(-1, +1)$	$-$
B	$(-1, +1)$	$(-1, -1)$	$+$
C	$(+1, -1)$	$(+1, -1)$	$+$
D	$(+1, +1)$	$(+1, +1)$	$-$



A/D & B/C are linearly separable

The middle would be the x-axis & margin would be distance between the points $(0,0)$ & $(0,-1)$ or $(0,0)$ and $(0,1)$.

$\boxed{\text{length of margin} = 1}$

Task 7: (10 points) Recall that the equation of the circle in the 2-dimensional plane is (x1 − a)^2 + (x2 − b)^2 − r2 = 0. Please expand out the formula and show that every circular region is linearly separable from the rest of the plane in the feature space (x1, x2, x1^2, x2^2).

HW-2.

$\boxed{7}$ Eq. : $(x_1-a)^2+(x_2-b)^2-r^2=0$

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

$$= \underbrace{(-r^2+a^2+b^2)}_{\boxed{1}} - 2a\underbrace{\boxed{x_1}}_{y_1} - 2b\underbrace{\boxed{x_2}}_{y_2} + \underbrace{\boxed{x_1^2}}_{y_3} + \underbrace{\boxed{x_2^2}}_{y_4} = 0$$

If we see each of the circled variables are linear functions, so they can be considered linearly seperable.

Task 8: (10 points) Recall that the equation of an ellipse in the 2-dimensional plane is c(x1 − a)^2 + d(x2 − b)^2 − 1 = 0. Please show that an SVM using the polynomial kernel of degree 2, K(u, v) = (1 + u · v)^2, is equivalent to a linear SVM in the feature space (1, x1, x2, x1^2, x2^2, x1x2) and hence that SVMs with this kernel can separate any elliptic region from the rest of the plane.

(8). Eq.: $c(x1-a)^2 + d(x2-b)^2 - 1 = 0$

$$= c(x_1^2 - 2ax_1 + a^2) + d(x_2^2 - 2bx_2 + b^2) - 1 = 0$$

$$= cx_1^2 - 2acx_1 + ca^2 + dx_2^2 - 2bdx_2 + db^2 - 1 = 0$$

$$= \underbrace{(ca^2 + db^2 - 1)}_{1} - \underbrace{2acx_1}_{x_1} - \underbrace{2bdx_2}_{x_2} + \underbrace{cx_1^2}_{x_1^2} + \underbrace{dx_2^2}_{x_2^2} = 0$$

Now we are given kernel

$$K(u,v) = (1+u.v)^2$$

$$\Rightarrow (1 + u_1v_1 + u_2v_2)^2$$

$$= 1 + 2u_1v_1 + 2u_2v_2 + u_1^2v_1^2 + u_2^2v_2^2 + 2u_1v_2v_1v_2.$$

$$\sim f(u).\, f(v)$$

where

$$f(u) = (1, u_1, u_2, u_1^2, u_2^2, u_1u_2)$$

& $f(v) = (1, v_1, v_2, v_1^2, v_2^2, v_1v_2).$

Hence with this kernel, eq. of elipse is linearly seperable in feature space of $(1, x1, x2, x1^2\ x2^2, x1x2)$.

Github Link: https://github.com/dpate243/CSE572-Data-Mining.git