PART-2

PROBLEM-2

PART-2

We were asked to use this DB & implement repeated two fold cross validation holdout.

We were also given the folds for each repitition too & the no. of repititions given are 4.

We need to use a_1 as the root.

We also considered the class C as nominal.

O	a_1	a_2	C
O_1	1	1	1
O_2	0	0	0
O_3	0	1	0
O_4	0	0	0
O_5	1	1	1
O_6	1	1	0
O_7	0	0	0
O_8	1	0	1

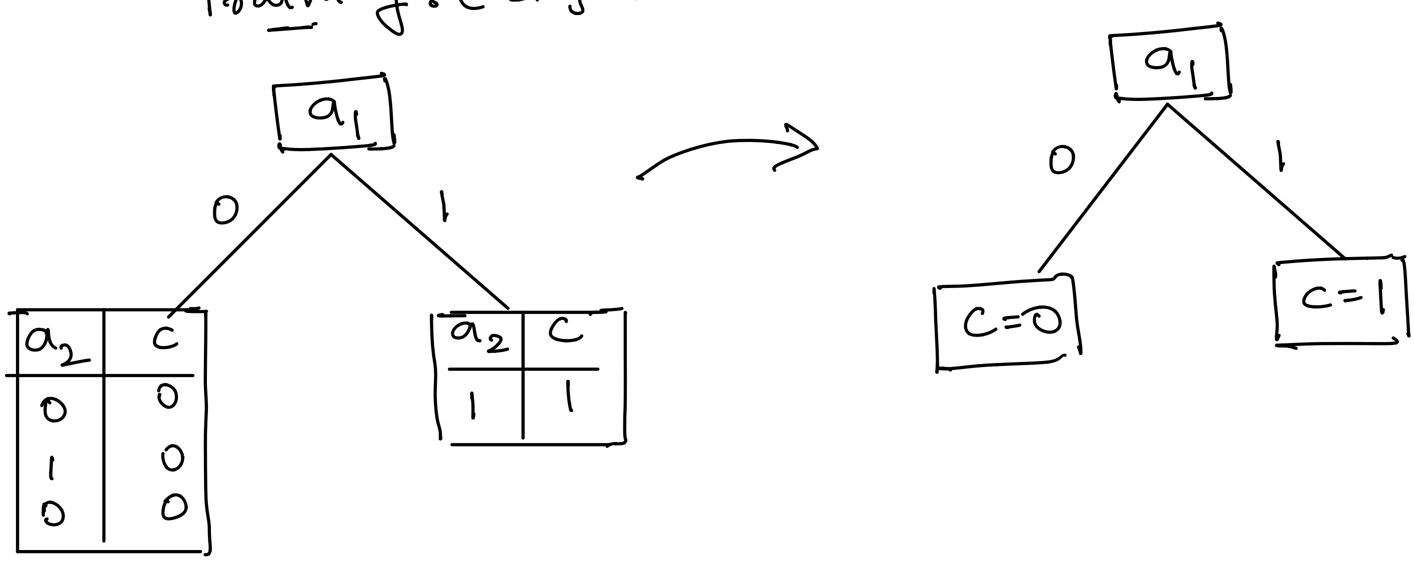
Step 1:

Repetition 1:

(1) train $= \{O_1, O_2, O_3, O_4\}$ test $= \{O_5, O_6, O_7, O_8\}$

So following the stages.

Training: (Stage 1)



As the decision tree is completed.

Discriminant rules :

Rule 1 : IF $a_1(x_1 = 0)$ THEN $c(x_1 = 0)$

Rule 2 : IF $a_1(x_1 = 1)$ THEN $c(x_1 = 1)$

We'll try to resubstitute as the part of stage 2.

train = $\{o_1, o_2, o_3, o_4\}$

The rule accuracy can be calculated as : 100%

o_1 is classified well using Rule 2.

o_2, o_3, o_4 are also classified well using Rule 1.

So stage 3 :

test : $\{o_5, o_6, o_7, o_8\}$.

o_5, o_8 are classified well using Rule 2.

o_7 is classified well using Rule 1.

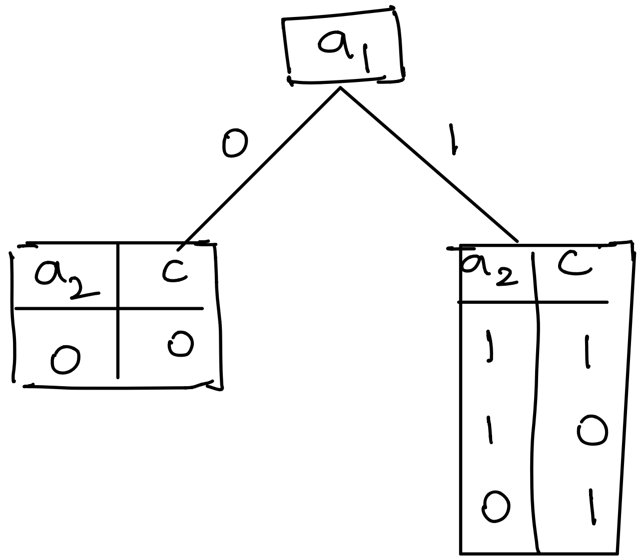
o_6 is misclassified from Rule 2.

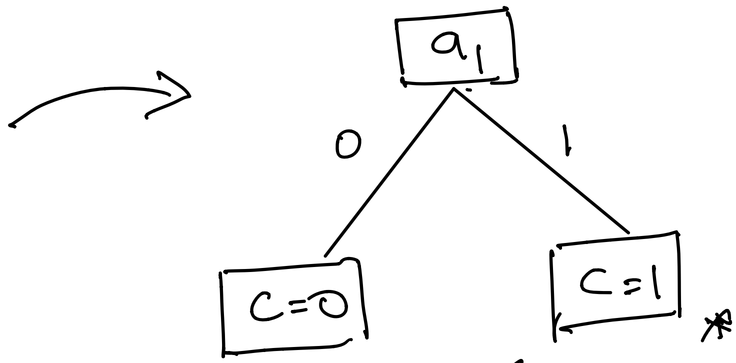
So the predictive accuracy would be $3/4 = 75\%$.

(2) Now for the second fold as training set.

train = $\{o_5, o_6, o_7, o_8\}$ test = $\{o_1, o_2, o_3, o_4\}$

So following the stages.





We pruned this tree, as there's no majority voting present when $a_2 = 1$ $\{1, 0\}$ where as for $a_2 = 0$ $\{1\}$. So when $a_2 = 1$, we went forward by keeping $c = 1$ as it is seen first.

As the decision tree is completed.

Discriminant rules:

Rule 1: IF $a_1(x_1 = 0)$ THEN $c(x_1 = 0)$

Rule 2: IF $a_1(x_1 = 1)$ THEN $c(x_1 = 1)$

We'll try to resubstitute as the part of Stage 2.

train = $\{o_5, o_6, o_7, o_8\}$

The rule accuracy can be calculated as : 3/4 : 75%

O_5 is well classified using Rule 2

O_6 is misclassified from Rule 2.

O_7 is well classified using Rule 1.

O_8 is well classified using Rule 2.

So stage 3:

test : $\{O_1, O_2, O_3, O_4\}$.

O_1 is classified well using Rule 2.

O_2, O_3, O_4 are classified well using Rule 1.

So the predictive accuracy would be 100%.

So our classifier would be the union of these rules. F_r.

Rule 1 : IF $a_1(x_i=0)$ THEN $c(x_i=0)$

Rule 2 : IF $a_1(x_i=1)$ THEN $c(x_i=1)$

Rule 3 : IF $a_1(x_i=0)$ THEN $c(x_i=0)$

Rule 4 : IF $a_1(x_i=1)$ THEN $c(x_i=1)$

After removing the repetitions.

It would be

Rule 1 : IF $a_1(x_i = 0)$ THEN $c(x_i = 0)$

Rule 2 : IF $a_1(x_i = 1)$ THEN $c(x_i = 1)$

The classifier's rule accuracy is $\frac{100 + 75}{2} = 87.5\%$

S_1 prediction accuracy is $\frac{100 + 75}{2} = 87.5\%$

Repetition 2 :

If we clearly check the given fold is

test = $\{o_5, o_6, o_7, o_8\}$, test = $\{o_1, o_2, o_3, o_4\}$

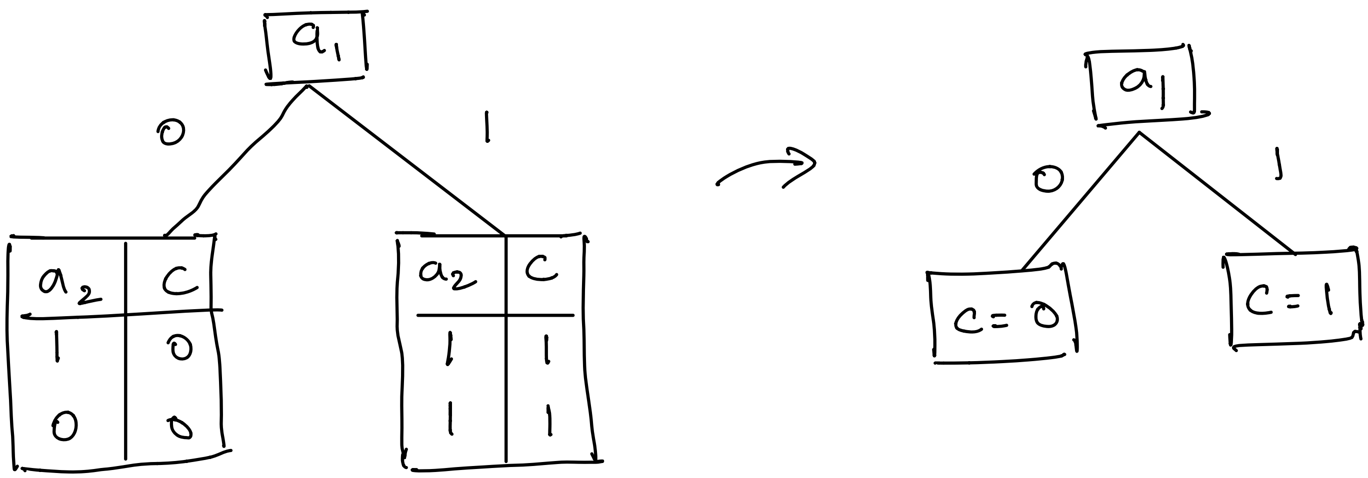
is exactly same as the 2^{nd} fold in the previous repetition. So, similarly the 2^{nd} fold in this case would be the same as the 1^{st} fold in previous scenario. So the classifier would be the same.

So $F_2 = F_1$

Repetition 3 :

(1) training = $\{o_1, o_3, o_5, o_7\}$ test = $\{o_2, o_4, o_6, o_8\}$.

Training : Stage 1 :



Discriminant rules:

Rule 1 : IF $a_1(x_1 = 0)$ THEN $c(x_1 = 0)$

Rule 2 : IF $a_1(x_1 = 1)$ THEN $c(x_1 = 1)$

So the rule accuracy is obtained by substituting the training data.

O_1 & O_5 are well classified using Rule 2.

O_3 & O_7 are well classified using Rule 1.

So the rule accuracy is 100%.

So during stage 3 for calculating predictive accuracy

O_2 & O_4 are well classified using Rule 1.

O_8 is well classified using Rule 2.

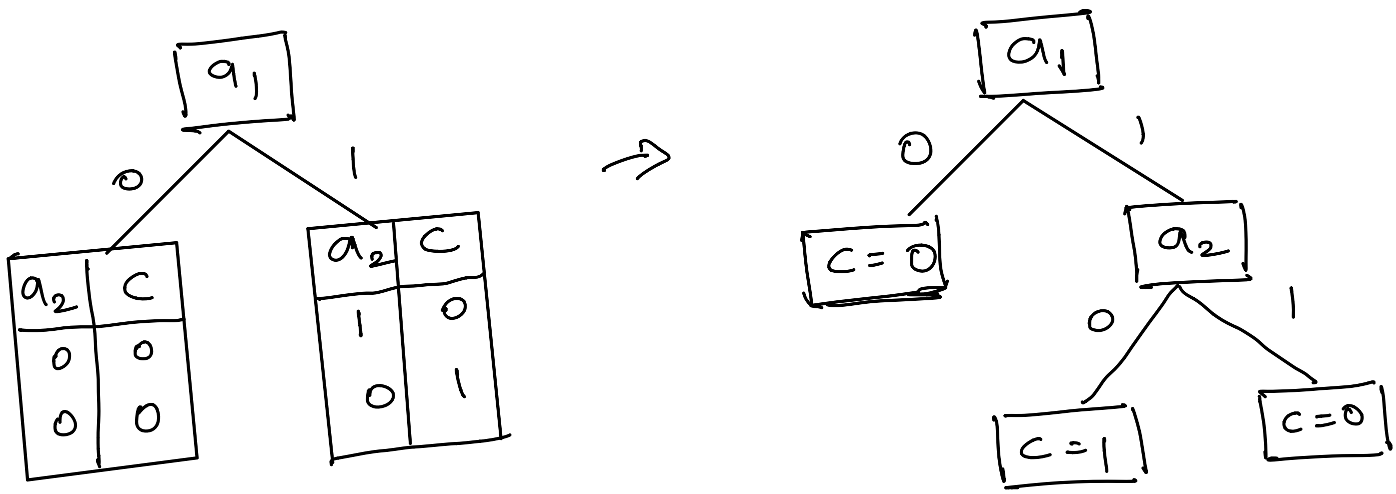
O_6 is misclassified from Rule 2.

So prediction accuracy = 75%.

(2)

training = $\{O_2, O_4, O_6, O_8\}$ test = $\{O_1, O_3, O_5, O_7\}$.

Training: Stage 1:



Discriminant Rules:

Rule 1: IF $a_1(x_1 = 0)$ THEN $c(x_1 = 0)$

Rule 2: IF $a_1(x_1 = 1)$ AND $a_2(x_1 = 0)$ THEN $c(x_1 = 1)$

Rule 3: IF $a_1(x_1 = 1)$ AND $a_2(x_1 = 1)$ THEN $c(x_1 = 0)$.

For stage 2 & resubstituting, the rule accuracy is

O_2 & O_4 are well classified using Rule 1.

O_6 is well classified using Rule 3.

O_8 is well classified using Rule 2.

So Rule Accuracy is 100%.

prediction Accuracy during stage 3 is

O_3 & O_7 are well classified using Rule 1.

O_1 & O_5 are misclassified from Rule 3.

So predictive accuracy is 50%.

So the combined rules of our classifier F_3 would be without repetitions.

Rule 1: IF $a_1(x_1 = 0)$ THEN $c(x_1 = 0)$

Rule 2: IF $a_1(x_1 = 1)$ AND $a_2(x_1 = 0)$ THEN $c(x_1 = 1)$

Rule 3: IF $a_1(x_1 = 1)$ AND $a_2(x_1 = 1)$ THEN $c(x_1 = 0)$.

Final Rule accuracy $= \frac{100 + 100}{2} = 100\%$.

Predictive accuracy would be $= \frac{50 + 75}{2}$

$= 62.5\%$.

Repetition 4: The folds same as repetition 3 & hence the classifier would be the same too

$F_4 = F_3$

Step 2:

After calculating the predictive & rule accuracies in the previous step.

We know that $F_1 = F_2$ & $F_3 = F_4$.

So let's compare F_1 & F_3.

	Rules Accuracy	Predictive Accuracy
F_1	87.5%	87.5%
F_3	100%	62.5%

Based on these details, Rules Accuracy is almost similar for both these classifiers but the predictive accuracy of F_3 is very low compared to F_1. So I would pick F_1 or F_2 as my classifier F. Let it be F_1

so $F_1 = F$.

Step 3 :

We need to construct the bagged ensemble classifier F^*.

F^* contains F_1, F_2, F_3 & F_4. Since $F_1 = F_2$ & $F_3 = F_4$. The majority voting is calculated just between F_1 & F_3.

Given data.

O	a_1	a_2
o_1	0	1
o_2	0	0
o_3	1	0
o_4	1	1

Since our F is F_1

the rules of this classifier are :

F :

Rule 1 : IF $a_1(x_1=0)$ THEN $c(x_1=0)$

Rule 2 : IF $a_1(x_1=1)$ THEN $c(x_1=1)$

And our F^* consists of F_1, F_2 $(F_1=F_2)$ & F_3, F_4 $(F_3=F_4)$

F^* : F_1

Rule 1 : IF $a_1(x_1=0)$ THEN $c(x_1=0)$

Rule 2 : IF $a_1(x_1=1)$ THEN $c(x_1=1)$

F_3

Rule 1: IF $a_1(x_1=0)$ THEN $c(x_1=0)$

Rule 2: IF $a_1(x_1=1)$ AND $a_2(x_1=0)$ THEN $c(x_1=1)$

Rule 3: IF $a_1(x_1=1)$ AND $a_2(x_1=1)$ THEN $c(x_1=0)$.

Let's first classify using F.

O_1 is classified 0 using Rule 1.

O_2 is classified 0 using Rule 1.

O_3 is classified 1 using Rule 2.

O_4 is classified 1 using Rule 2.

for F^*

F_1 is classified similar to F as above.

Where as F_3.

O_1 is classified 0 using Rule 1.

O_2 is classified 0 using Rule 1.

O_3 is classified 1 using Rule 2.

O_4 is classified 0 using Rule 3.

Since we need to take majority vote.

of F_1 & F_3 for F^*. The final result would be.

O_1 & O_2 are classified as 0.

O_3 & O_4 are classified as 1.

(O_4 doesn't have a majority vote, so we went with 1 as it's seen first).

	F	F^*
O_1	0	0
O_2	0	0
O_3	1	1
O_4	1	1

So based on our results, both these classifiers provided the same results.