22nd February, 2024

BRAC UNIVERSITY
Department of Computer Science and Engineering
B.Sc. in Computer Science / CSE Program
Quiz-1, Spring 2024

Course:	**CSE437 (Data Science: Coding With Real World Data), Section-1**
Instructor:	**Dr. Md. Golam Rabiul Alam**
Full Marks:	**10**
Time:	**60 minutes**

Note: Course Outcome (CO), Cognitive Level, and Mark of each question are mentioned at the right margin.

---

1. Consider building an ensemble of decision stumps G_m with the AdaBoost algorithm. [CO3, C3, Marks:7+3]

$$f(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

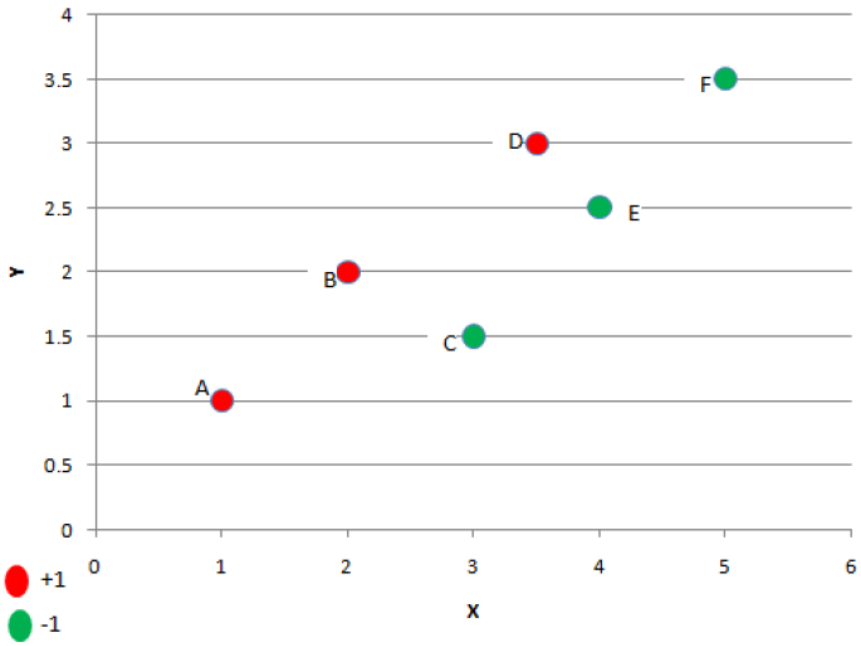


Figure: 1 Dataset. The points A(1,1), B(2, 2), D(3.5, 3) are in positive(+1) class and the points C(3, 1.5), E(4, 2.5), F(5, 3.5) are in negative(-1) class.

Determine the first decision stump and draw both the decision boundary and its positive orientation. Mention all the point(s) whose weight will increase as a result of incorporating the first stump (the weight update due to the first stump). **Prepare** a new dataset based on the first stump considering the following uniformly distributed random numbers:
[0.15, 0.30, 0.45, 0.60, 0.75, 0.90]

Point	*X*	*Y*	*Class*
A	1	1	positive
B	2	2	positive
C	3	1.5	negative
D	3.5	3	positive
E	4	2.5	negative
F	5	3.5	negative

	X	Class		Number of Left side instances with "Positive class"	Number of Left side instances with "Negative"	GINI score of Left side instances	Number of Right side instances with "Positive"	Number of Right side instances with "Negative Class"	GINI score of Right side instances	GINI score of Weight
A	1	positive	**Considering X< 1.5 as the separator**	1	0	0	2	3	**0.48**	0.4
B	2	positive	**Considering X<2.5 as the separator**	2	0	0	1	3	**0.375**	0.25
C	3	negative	**Considering X<3.25 as the separator**	2	1	0.444444	1	2	**0.444444444**	0.444444
D	3.5	positive	**Considering X<3.75 as the separator**	3	1	0.375	0	2	**0**	0.25
E	4	negative	**Considering X<4.5 as the separator**	3	2	0.48	0	1	**0**	0.4
F	5	negative								
									Minimum GINI=	0.25

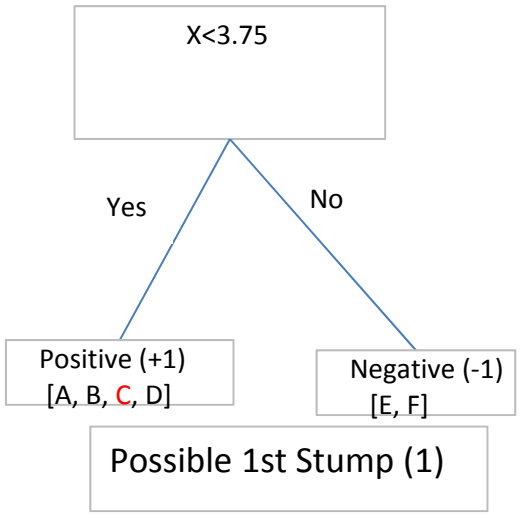
GINI Impurity of X (X<2.5) or (X<3.75)=	0.25

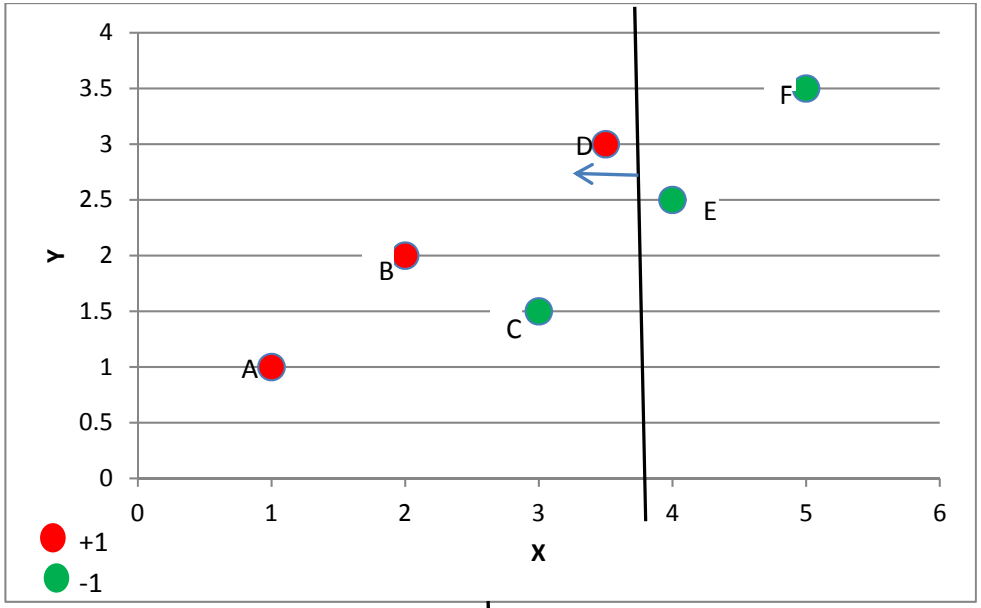
	Y	Class		Down side instances	Down side instances	score of Left side	Up side instances	side instances with "Negative	Right side instances	score of Weight
A	1	positive	**Considering Y< 1.25 as the separator**	1	0	0	2	3	**0.48**	0.4
C	1.5	negative	**Considering Y<1.75 as the separator**	1	1	0.5	2	2	**0.5**	0.5
B	2	positive	**Considering Y<2.25 as the separator**	2	1	0.44444	1	2	**0.444444444**	0.44444
E	2.5	negative	**Considering Y<2.75 as the separator**	2	2	0.5	1	1	**0.5**	0.5
D	3	positive	**Considering Y<3.25 as the separator**	3	2	0.48	0	1	**0**	0.4
F	3.5	negative								
									Minimum GINI=	0.4

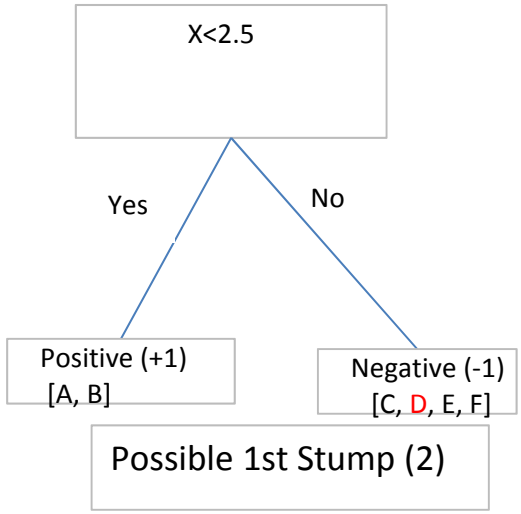
GINI Impurity of Y (Y<1.25) or (Y<3.25)= 0.4

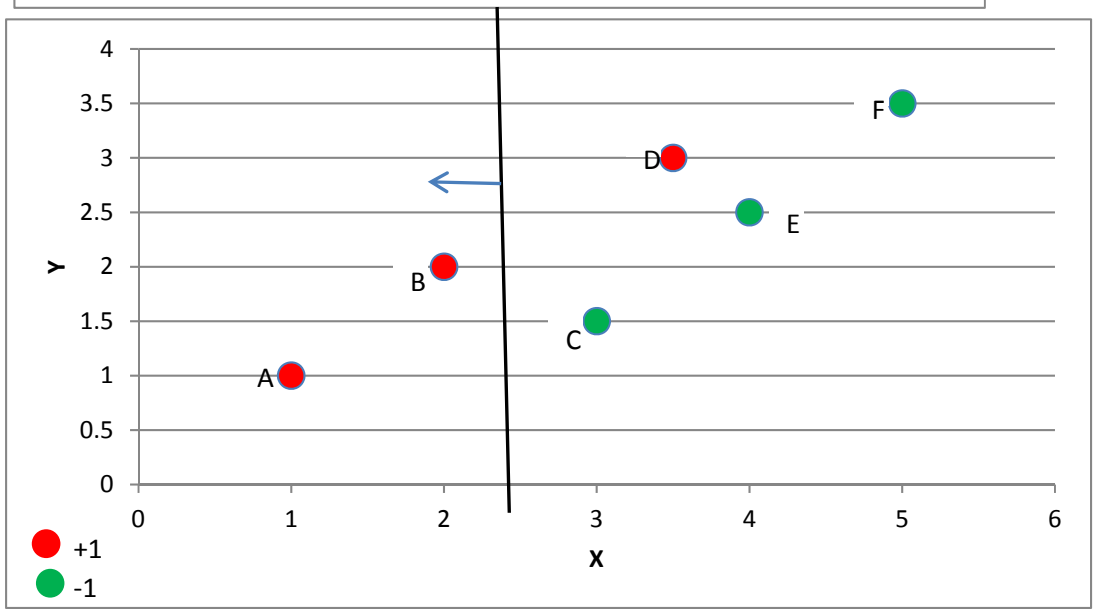
GINI Impurity of X (X<2.5) or (X<3.75)= 0.25

So, 1st Splitters are= (X<2.5) or (X<3.75)


X<3.75
Yes
No
Positive (+1)
[A, B, C, D]
Negative (-1)
[E, F]
Possible 1st Stump (1)


4
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
y
x
A
B
C
D
E
F
+1
-1


X<2.5
Yes
No
Positive (+1)
[A, B]
Negative (-1)
[C, D, E, F]
Possible 1st Stump (2)


4
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
y
x
A
B
C
D
E
F
+1
-1

Using the 1st Spliter (X<3.75)

Point	*X*	*Y*	*Class*	Sample Weight
A	1	1	positive	1/6
B	2	2	positive	1/6
C	3	1.5	negative	1/6
D	3.5	3	positive	1/6
E	4	2.5	negative	1/6
F	5	3.5	negative	1/6

	Total Error=	1*1/6+0*1/6	0.166666667
Amount of say =	α1	(1/2)*log((1-total error)/total error)	0.804718956

Point	*X*	*Y*	*Class*	Sample Weight	*Updated Weight Formula*	*Updated Weight*	Normalized Weight	Cumulative Normalized Sample Weight	Generated Random Numbers	
A	1	1	positive	1/6	1/6*e^(-α1)	0.074536	0.1	0.1	0.15	
B	2	2	positive	1/6	1/6*e^(-α1)	0.074536	0.1	0.2	0.3	
C	3	1.5	negative	1/6	1/6*e^(+α1)	0.372678	0.5	0.7	0.45	
D	3.5	3	positive	1/6	1/6*e^(-α1)	0.074536	0.1	*0.8*	*0.6*	
E	4	2.5	negative	1/6	1/6*e^(-α1)	0.074536	0.1	0.9	0.75	
F	5	3.5	negative	1/6	1/6*e^(-α1)	0.074536	0.1	1	0.9	
					Total=	0.745356				

The Derived New DataSet

Point	*X*	*Y*	*Class*
B	2	2	positive
C	3	1.5	negative
C	3	1.5	negative
C	3	1.5	negative
D	3.5	3	positive
E	4	2.5	negative

	X	**Class**		**Number of Left side instances with "Positive class"**	**Number of Left side instances with "Negative"**	**GINI score of Left side instances**	**Number of Right side instances with "Positive"**	**Number of Right side instances with "Negative Class"**	**GINI score of Right side instances**	**GINI score of Weight**
B	2	positive	**Considering X< 2.5 as the separator**	1	0	0	1	4	**0.32**	0.266667
C	3	negative	**Considering X<=3 as the separator**	1	3	0.375	1	1	**0.5**	0.416667
C	3	negative	**Considering X<=3 as the separator**	1	3	0.375	1	1	**0.5**	0.416667
C	3	negative	**Considering X<3.25 as the separator**	1	3	0.375	1	1	**0.5**	0.416667
D	3.5	positive	**Considering X<3.75 as the separator**	2	3	0.48	0	1	**0**	0.4
E	4	negative								
									Minimum GINI=	0.266667

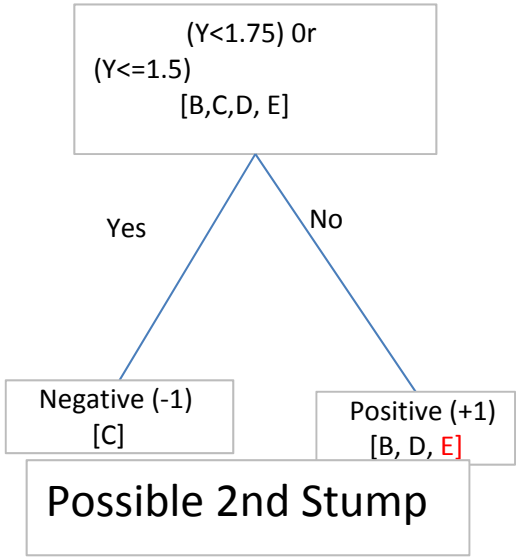
GINI Impurity of X (X<2.5) or (X<3.0)= 0.266666667

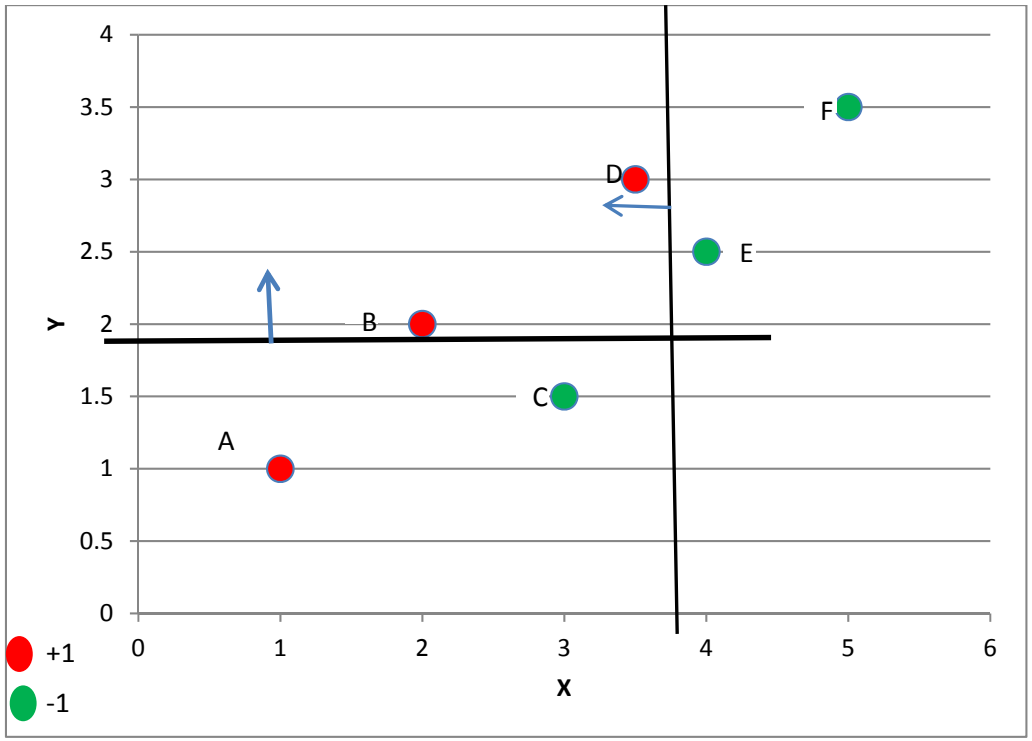
	Y	Class		Number of Down side instances with "Positive class"	Number of Down side instances with "Negative"	GINI score of Left side instances	Number of Up side instances with "Positive"	Number of Up side instances with "Negative Class"	GINI score of Right side instances	GINI score of Weight
C	1.5	negative	**Considering Y<= 1.5 as the separator**	0	3	0	2	1	**0.444444444**	0.222222
C	1.5	negative	**Considering Y<=1.5 as the separator**	0	3	0	2	1	**0.444444444**	0.222222
C	1.5	negative	**Considering Y<1.75 as the separator**	0	3	0	2	1	**0.444444444**	0.222222
B	2	positive	**Considering Y<2.25 as the separator**	1	3	0.375	1	1	**0.5**	0.416667
E	2.5	negative	**Considering Y<2.75 as the separator**	1	4	0.32	1	0	**0**	0.266667
D	3	positive								
									Minimum GINI=	0.222222

GINI Impurity of Y (Y<=1.5) or (Y<1.75)= 0.222222222

GINI Impurity of X (X<2.5) or (X<3.75)= 0.266666667

So, 2nd Splitters are= (Y<=1.5) or (Y<1.75)





Using the 1st Spliter (X<2.5)

Point	*X*	*Y*	*Class*	Sample Weight
A	1	1	positive	1/6
B	2	2	positive	1/6
C	3	1.5	negative	1/6
D	3.5	3	positive	1/6
E	4	2.5	negative	1/6
F	5	3.5	negative	1/6

	Total Error=	0*1/6+1*1/6	0.166666667
Amount of say =	α1	(1/2)*log((1-total error)/total error)	0.804718956

Point	*X*	*Y*	*Class*	Sample Weight	*Updated Weight Formula*	*Updated Weight*	Normalized Weight	Cumulative Normalized Sample Weight	Generated Random Numbers
A	1	1	positive	1/6	1/6*e^(-α1)	0.074536	0.1	0.1	0.15
B	2	2	positive	1/6	1/6*e^(-α1)	0.074536	0.1	0.2	0.3
C	3	1.5	negative	1/6	1/6*e^(+α1)	0.074536	0.1	0.3	0.45
D	3.5	3	positive	1/6	1/6*e^(-α1)	0.372678	0.5	*0.8*	*0.6*
E	4	2.5	negative	1/6	1/6*e^(-α1)	0.074536	0.1	0.9	0.75
F	5	3.5	negative	1/6	1/6*e^(-α1)	0.074536	0.1	1	0.9
					Total=	0.745356			

The Derived New DataSet

Point	*X*	*Y*	*Class*
B	2	2	positive
C	3	1.5	negative
D	3.5	3	positive
D	3.5	3	positive
D	3.5	3	positive
E	4	2.5	negative

	X	Class		Number of Left side instances with "Positive class"	Number of Left side instances with "Negative"	GINI score of Left side instances	Number of Right side instances with "Positive"	Number of Right side instances with "Negative Class"	GINI score of Right side instances	GINI score of Weight
			Considering X< 2.5 as the separator	1	0	0	3	2	**0.48**	
B	2	positive								0.4
			Considering X<3.25 as the separator	1	1	0.5	3	1	**0.375**	
C	3	negative								0.416667
			Considering X<=3.5 as the separator	4	1	0.32	0	1	**0**	
D	3.5	positive								0.266667
			Considering X<=3.5 as the separator	4	1	0.32	0	1	**0**	
D	3.5	positive								0.266667
			Considering X<3.75 as the separator	4	1	0.32	0	1	**0**	
D	3.5	positive								0.266667
E	4	negative								
									Minimum GINI=	0.266667

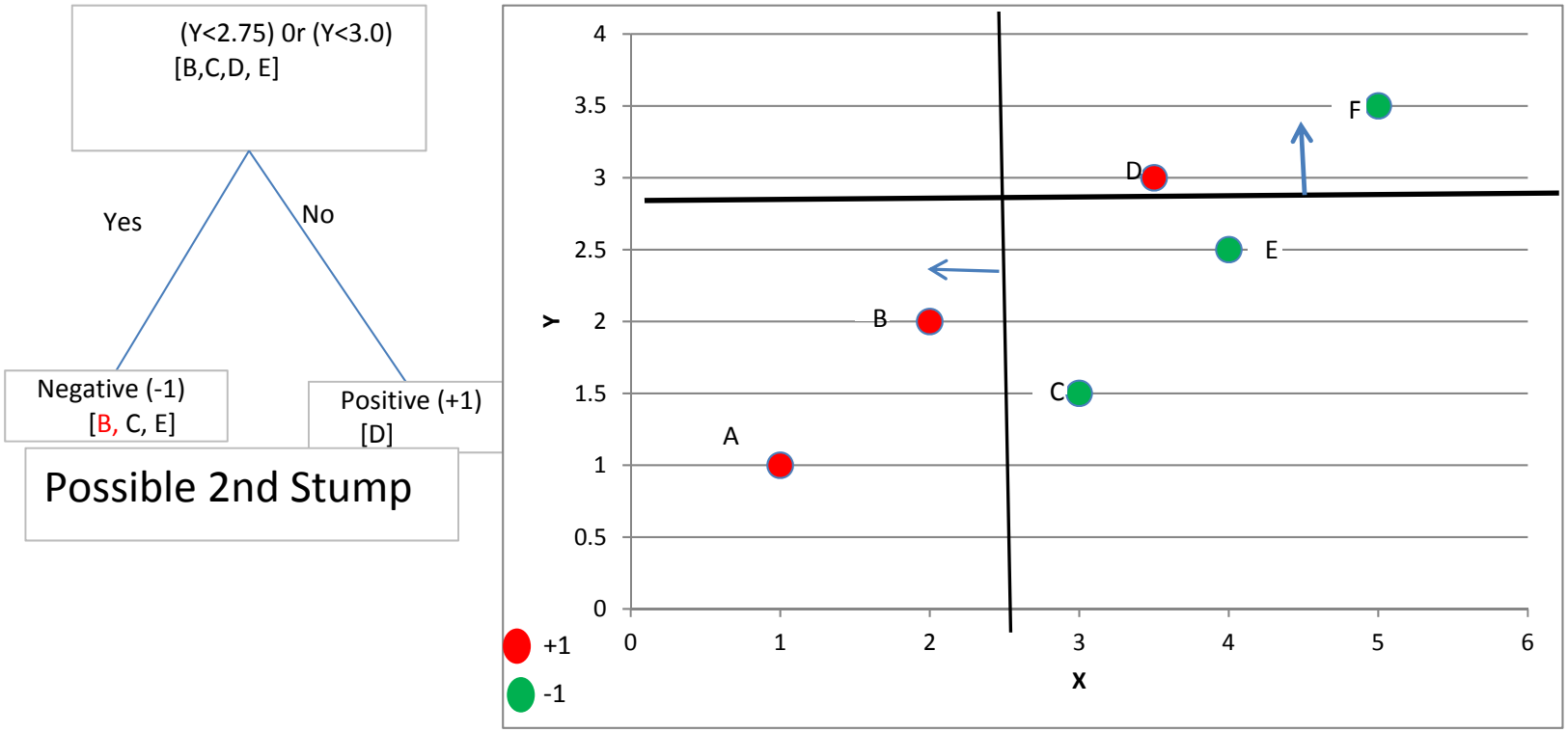
GINI Impurity of X (X<=3.5) or (X<3.75)= 0.26666667

	Y	Class		Number of Down side instances with "Positive class"	Number of Down side instances with "Negative"	GINI score of Left side instances	Number of Up side instances with "Positive"	Number of Up side instances with "Negative Class"	GINI score of Right side instances	GINI score of Weight
C	1.5	negative	**Considering Y< 1.75 as the separator**	0	1	0	4	1	**0.32**	0.266667
B	2	positive	**Considering Y<2.25 as the separator**	1	1	0.5	3	1	**0.375**	0.416667
E	2.5	negative	**Considering Y<2.75 as the separator**	1	2	0.4444	3	0	**0**	0.222222
D	3	positive	**Considering Y<3.0 as the separator**	1	2	0.4444	3	0	**0**	0.222222
D	3	positive	**Considering Y<3.0 as the separator**	1	2	0.4444	3	0	**0**	0.222222
D	3	positive								
									Minimum GINI=	0.222222

GINI Impurity of Y (Y<2.75) or (Y<3.0)= 0.22222222

So, 2nd Splitters are= (Y<2.75) or (Y<3.0)

GINI Impurity of X (X<2.5) or (X<3.75)= 0.26666667


(Y<2.75) 0r (Y<3.0)
[B,C,D, E]
Yes
No
Negative (-1)
[B, C, E]
Positive (+1)
[D]
Possible 2nd Stump
4
3.5
3
2.5
2
1.5
1
0.5
0
Y
0
1
2
3
4
5
6
X
A
B
C
D
E
F
+1
-1