DATA 301: Decision Trees
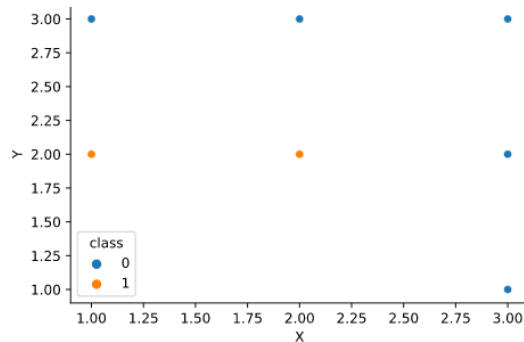## A Supervised Algorithm

Topics

Demonstration
Impurity
Choosing which feature to split on
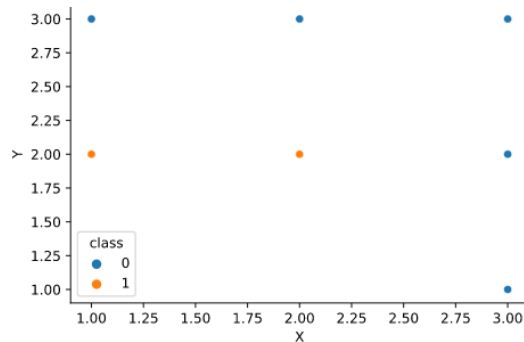scikit-learn

Problem

What if data is not linearly separable?

For this data you cannot draw a line that cleanly separates class 0 from class 1
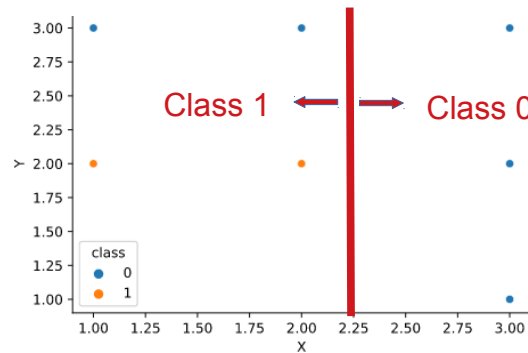
Problem

What if data is not linearly separable?

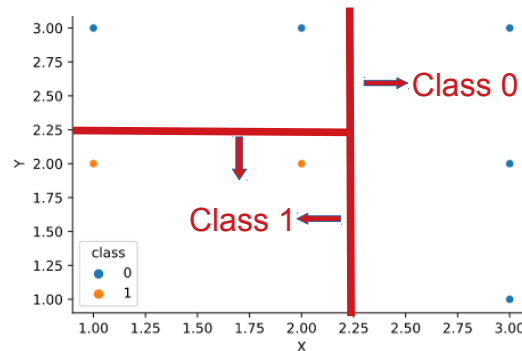But you can use multiple line segments to do so.

Problem

What if data is not linearly separable?

But you can use multiple line segments to do so.
The first

Problem

What if data is not linearly separable?

But you can use multiple line segments to do so.
The first
The second

Problem

Easy to do for 2 dimensions, what about 4 or 8 dimensions?
How to do this algorithmically?
Especially since data looks like this.

	X	Y	class
0	1	3	a
1	2	3	a
2	3	3	a
3	1	2	b
4	2	2	b
5	3	2	a
6	3	1	a

Solution

Using 2 common algorithms:

- entropy and information gain
- gini impurity

We will use gini impurity for this class. It's calculated with the following equation

gini_impurity=1- (probability class A)**2 - (probability class B)**2
          - (probability class C)**2 … (probability class n)**2

Solution

	X	Y	class
0	1	3	a
1	2	3	a
2	3	3	a
3	1	2	b
4	2	2	b
5	3	2	a
6	3	1	a

Back to the problem
We want to predict class using features X and Y

Where to start?

Solution

	X	Y	class
0	1	3	a
1	2	3	a
2	3	3	a
3	1	2	b
4	2	2	b
5	3	2	a
6	3	1	a

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
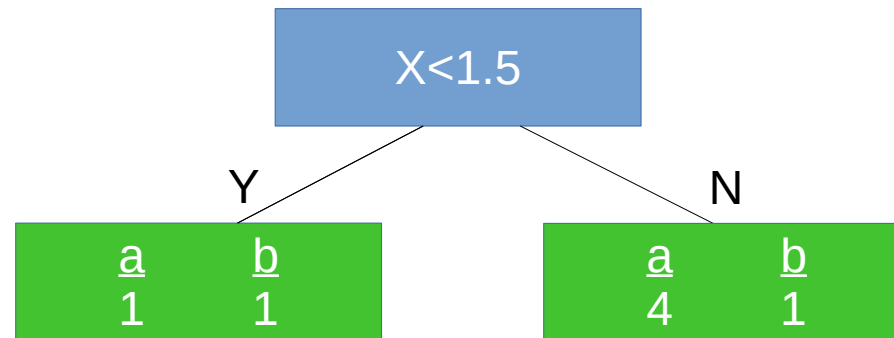Find gini impurity using each of these midpoints

Solution

gini_impurity=1- (probability class A)**2 - (probability class B)**2

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

X<1.5

Y   N

a	b
1	1

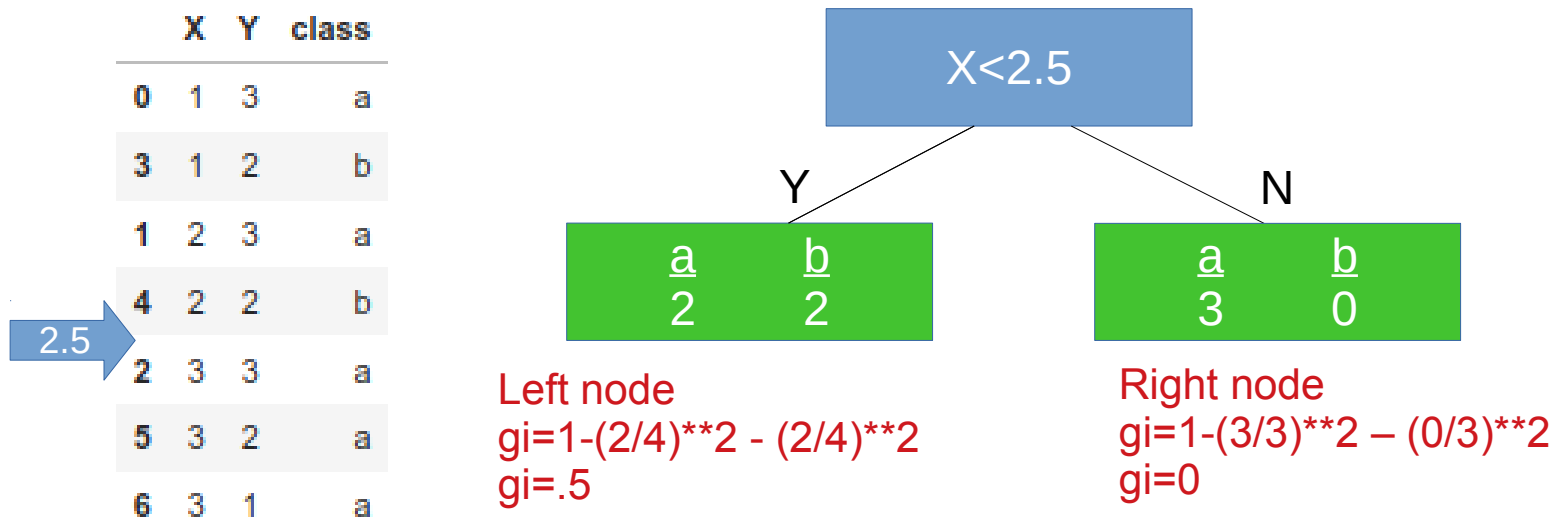| a	b
4	1

Left node
gi=1-(1/2)**2 - (1/2)**2
gi=.5

Right node
gi=1-(4/5)**2 – (1/5)**2
gi=.32

Gini impurity for column X at split 1.5 is a weighted average though
Weights = (total values in node)/(total values in both nodes)

GI= 2/(2+5)*.5 + 5/(2+5)*.32
GI=.37

Solution

gini_impurity=1- (probability class A)**2 - (probability class B)**2

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

2.5

X<2.5

Y   N

a	b
2	2

a	b
3	0

Left node
gi=1-(2/4)**2 - (2/4)**2
gi=.5

Right node
gi=1-(3/3)**2 – (0/3)**2
gi=0

Gini impurity for column X at split 2.5 is a weighted average though
Weights = (total values in node)/(total values in both nodes)

GI= 4/(2+5)*.5 + 3/(2+5)*0
GI=.28

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
    2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
   1.5 => GI=.37
    2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5 →

2.5 →

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
     2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Repeat for all other columns (except the target 'class')

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

Back to the problem
We want to predict class using features X and Y

Where to start?

Choose column (Y)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
  1.5 => GI=.37
  2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Repeat for all other columns (except the target 'class')

Solution

	X	Y	class
6	3	1	a
3	1	2	b
4	2	2	b
5	3	2	a
0	1	3	a
1	2	3	a
2	3	3	a

Back to the problem
We want to predict class using features X and Y

Where to start?

Choose column (Y)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
    2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Repeat for all other columns (except the target 'class')

Solution

	X	Y	class
6	3	1	a
3	1	2	b
4	2	2	b
5	3	2	a
0	1	3	a
1	2	3	a
2	3	3	a

Back to the problem
We want to predict class using features X and Y

Where to start?

Choose column (Y)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
     2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Repeat for all other columns (except the target 'class')

Solution

	X	Y	class
6	3	1	a
3	1	2	b
4	2	2	b
5	3	2	a
0	1	3	a
1	2	3	a
2	3	3	a

1.5

2.5

Back to the problem
We want to predict class using features X and Y

Where to start?

Choose column (Y)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
  1.5 => GI=.37
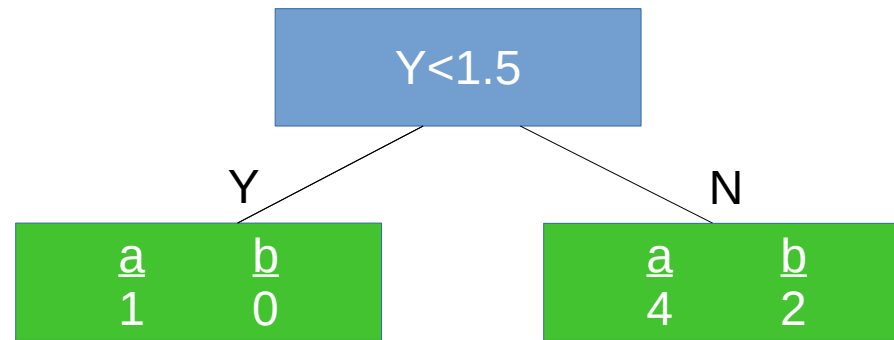   2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Repeat for all other columns (except the target 'class')

Solution

gini_impurity=1- (probability class A)**2 - (probability class B)**2

	X	Y	class
6	3	1	a
3	1	2	b
4	2	2	b
5	3	2	a
0	1	3	a
1	2	3	a
2	3	3	a

1.5

Y<1.5

Y

N

a	b
1	0

a	b
4	2

Left node
gi=1-(1/1)**2 - (0/1)**2
gi=0

Right node
gi=1-(4/6)**2 – (2/6)**2
gi=.44

Gini impurity for column Y at split 2.5 is a weighted average though
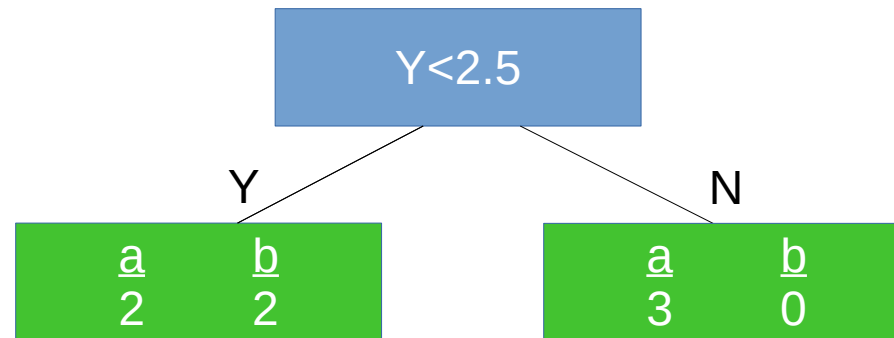Weights = (total values in node)/(total values in both nodes)

GI= 1/(1+6)*0 + 6/(1+6)*.44
GI=.37

Solution

gini_impurity=1- (probability class A)**2 - (probability class B)**2

	X	Y	class
6	3	1	a
3	1	2	b
4	2	2	b
5	3	2	a
0	1	3	a
1	2	3	a
2	3	3	a

2.5

Y<2.5

Y   N

a	b
2	2

a	b
3	0

Left node
gi=1-(2/4)**2 - (2/4)**2
gi=.5

Right node
gi=1-(3/3)**2 – (0/3)**2
gi=0

Gini impurity for column X at split 2.5 is a weighted average though
Weights = (total values in node)/(total values in both nodes)

GI= 4/(2+5)*.5 + 3/(2+5)*0
GI=.28

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
     2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Col Y→ Choose split 2.5 with GI=.28

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5 →
2.5 →

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
    2.5=> GI=.28
Choose split with lowest value
Col X → Choose split 2.5 with GI=.28
Col Y→ Choose split 2.5 with GI=.28
Why are they both the same?

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5 →

2.5 →

Back to the problem
We want to predict class using features X and Y

Where to start?

Start with first column (X)
Sort it ascending
Find midpoints between adjacent values
Find gini impurity using each of these midpoints
    1.5 => GI=.37
     2.5=> GI=.28
Choose split with lowest value
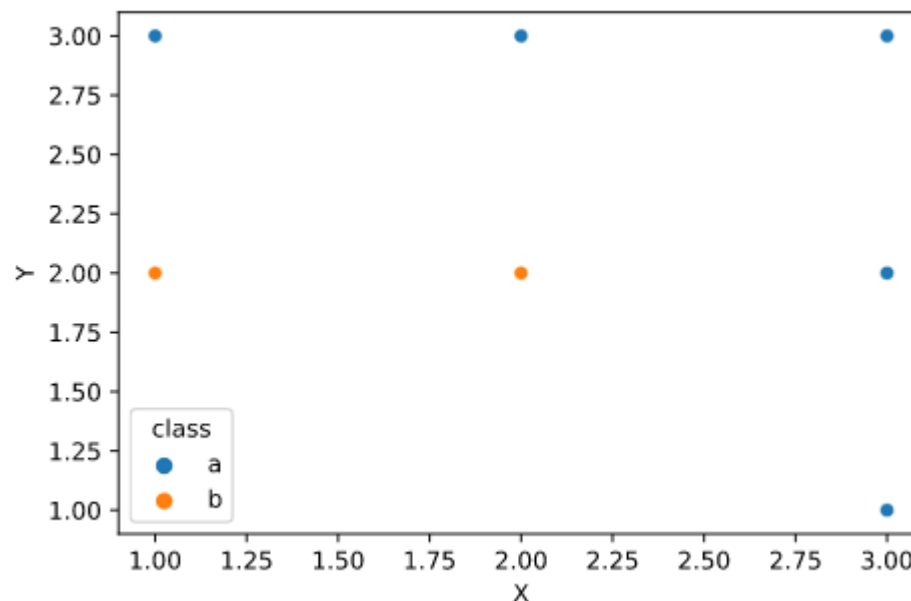Col X → Choose split 2.5 with GI=.28
Col Y→ Choose split 2.5 with GI=.28
Why are they both the same?

Solution

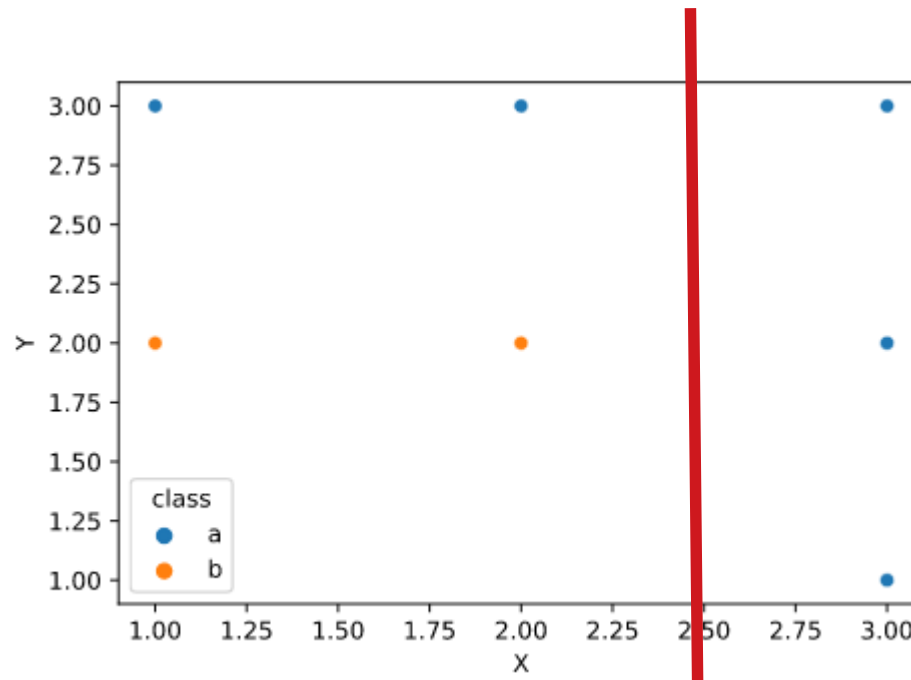|	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Why are they both the same?
Because it doesn't matter whether we split on
X=2.5 or Y=2.5,
we still have 3 a's on 1 side and 2 a's and 2 b's on the other

Solution

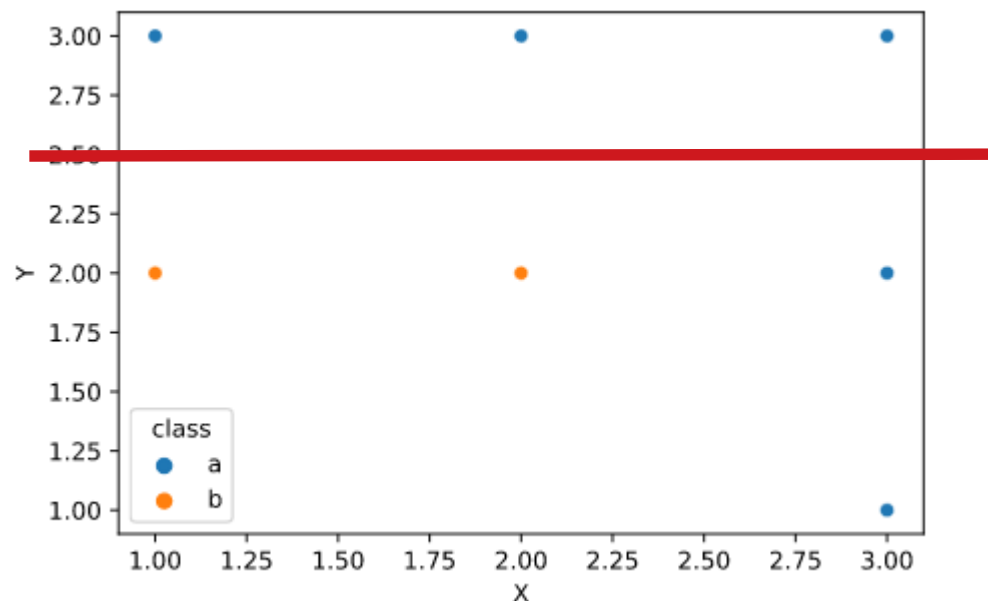|	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Why are they both the same?
Because it doesn't matter whether we split on
X=2.5 or Y=2.5,
we still have 3 a's on 1 side and 2 a's and 2 b's on the other

Solution

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

Why are they both the same?
Because it doesn't matter whether we split on
X=2.5 or Y=2.5,
we still have 3 a's on 1 side and 2 a's and 2 b's on the other
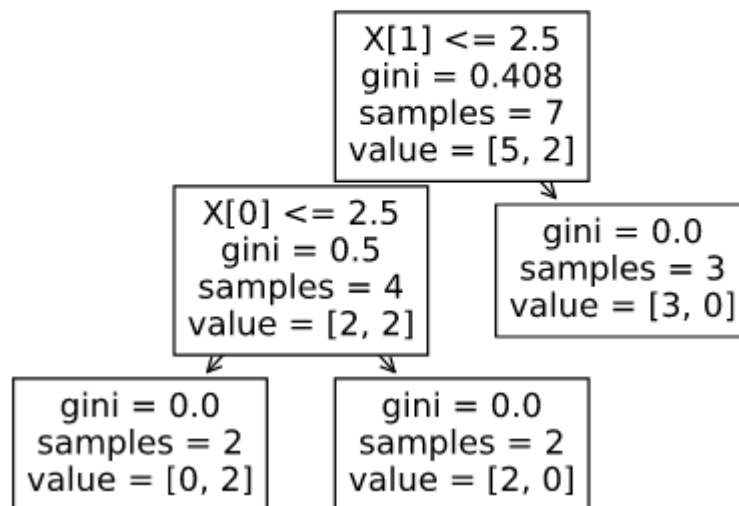
Scikitlearns Decision Tree

	X	Y	class
0	1	3	a
3	1	2	b
1	2	3	a
4	2	2	b
2	3	3	a
5	3	2	a
6	3	1	a

1.5

2.5

```
from sklearn import tree
df2=df.to_numpy();
clf = tree.DecisionTreeClassifier()
clf.fit(X=df2[:,:-1],y=df2[:,-1]);
```

```
tree.plot_tree(clf);
```

X[1] <= 2.5
gini = 0.408
samples = 7
value = [5, 2]

X[0] <= 2.5
gini = 0.5
samples = 4
value = [2, 2]

gini = 0.0
samples = 3
value = [3, 0]

gini = 0.0
samples = 2
value = [0, 2]

gini = 0.0
samples = 2
value = [2, 0]

Summary

How to split by eye
What gini impurity is
How to use gini impurity to find the best features to split on and the best split points
Scikit learns decision tree