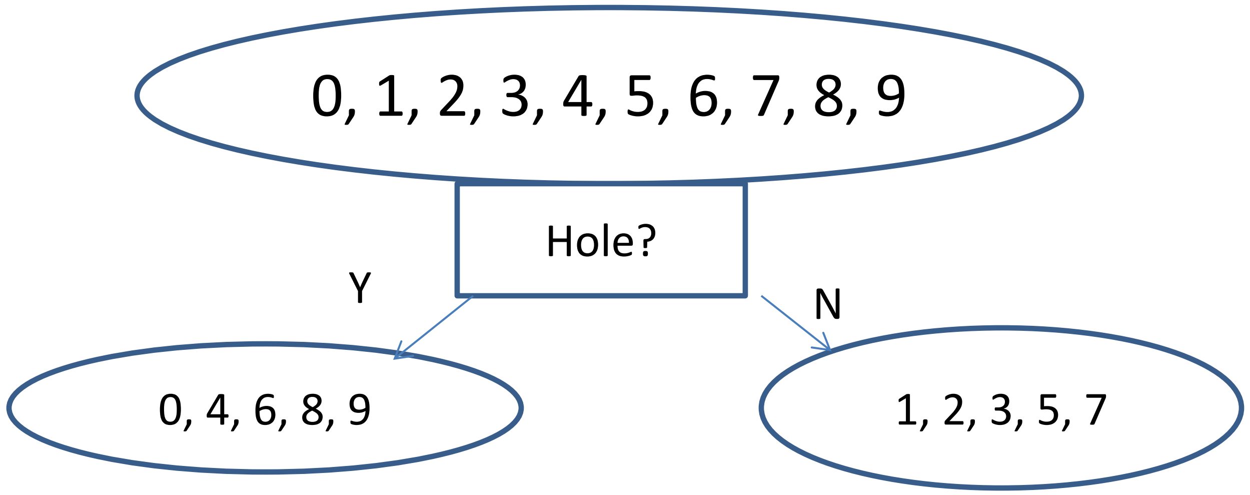
Decision tree

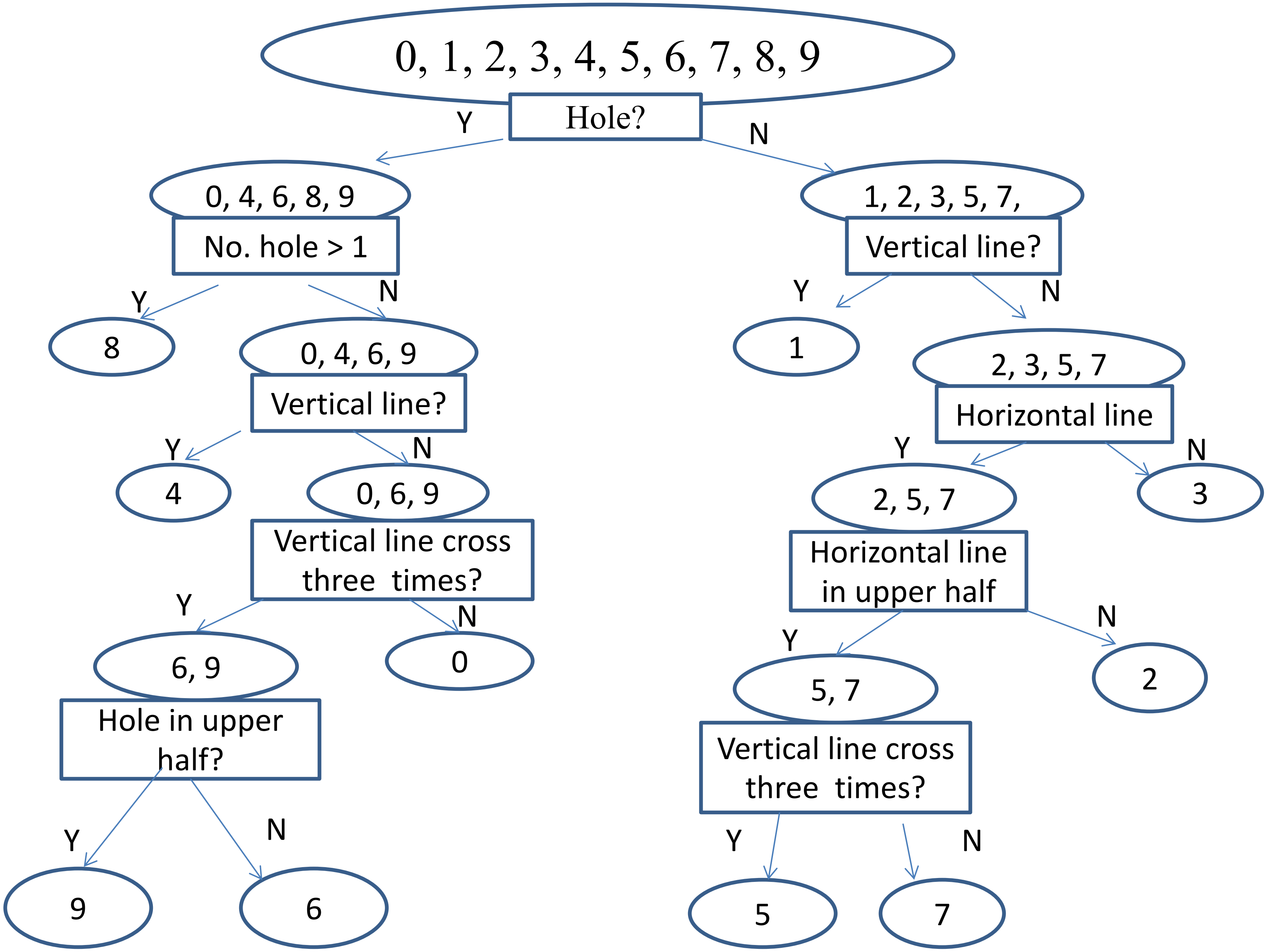
Motivating example: postal letter sorting

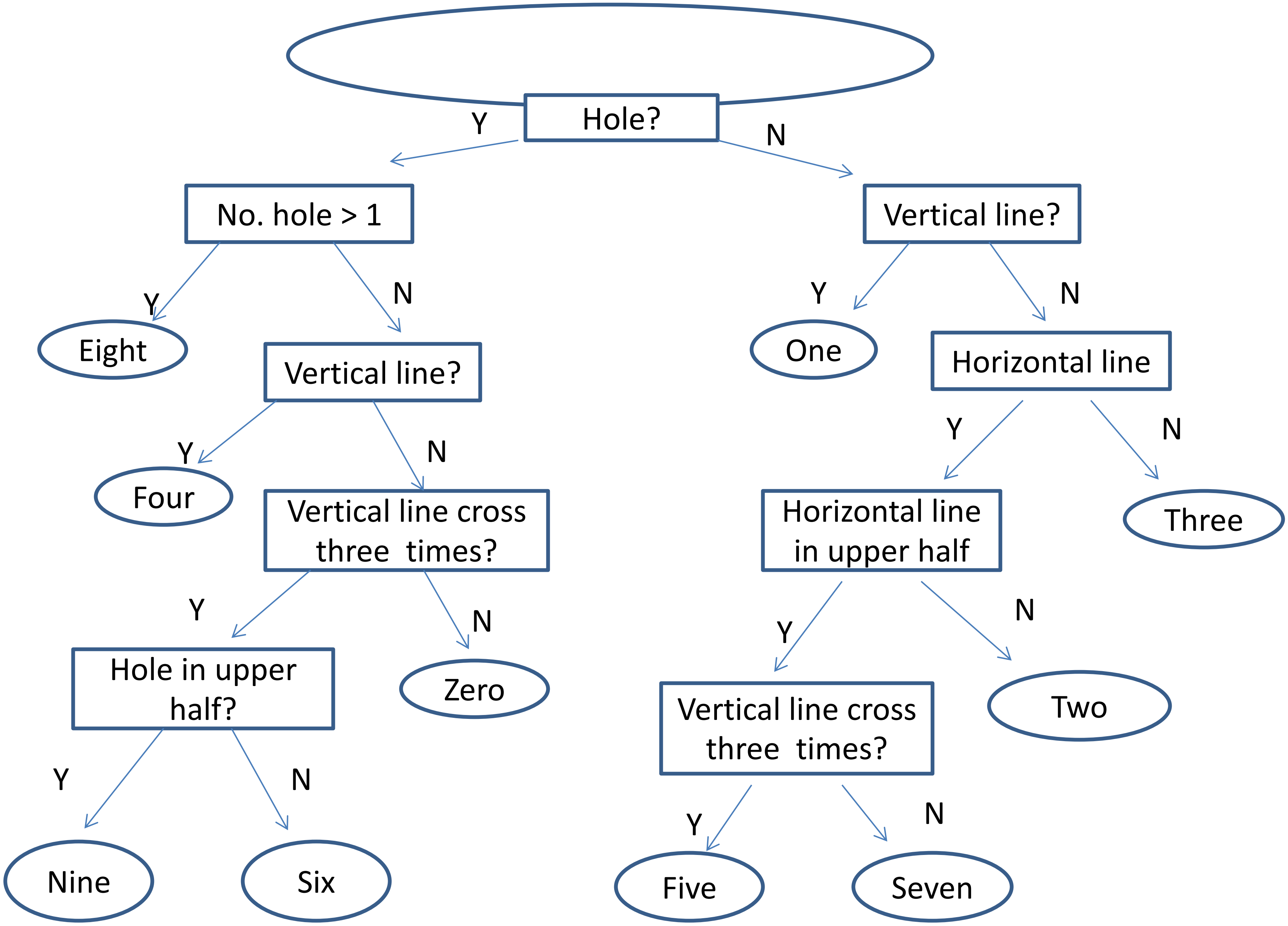
- Sorting postal letter according to pin number.
- To do that you have to recognize each numerical digits(0, 1, 2, … , 9).

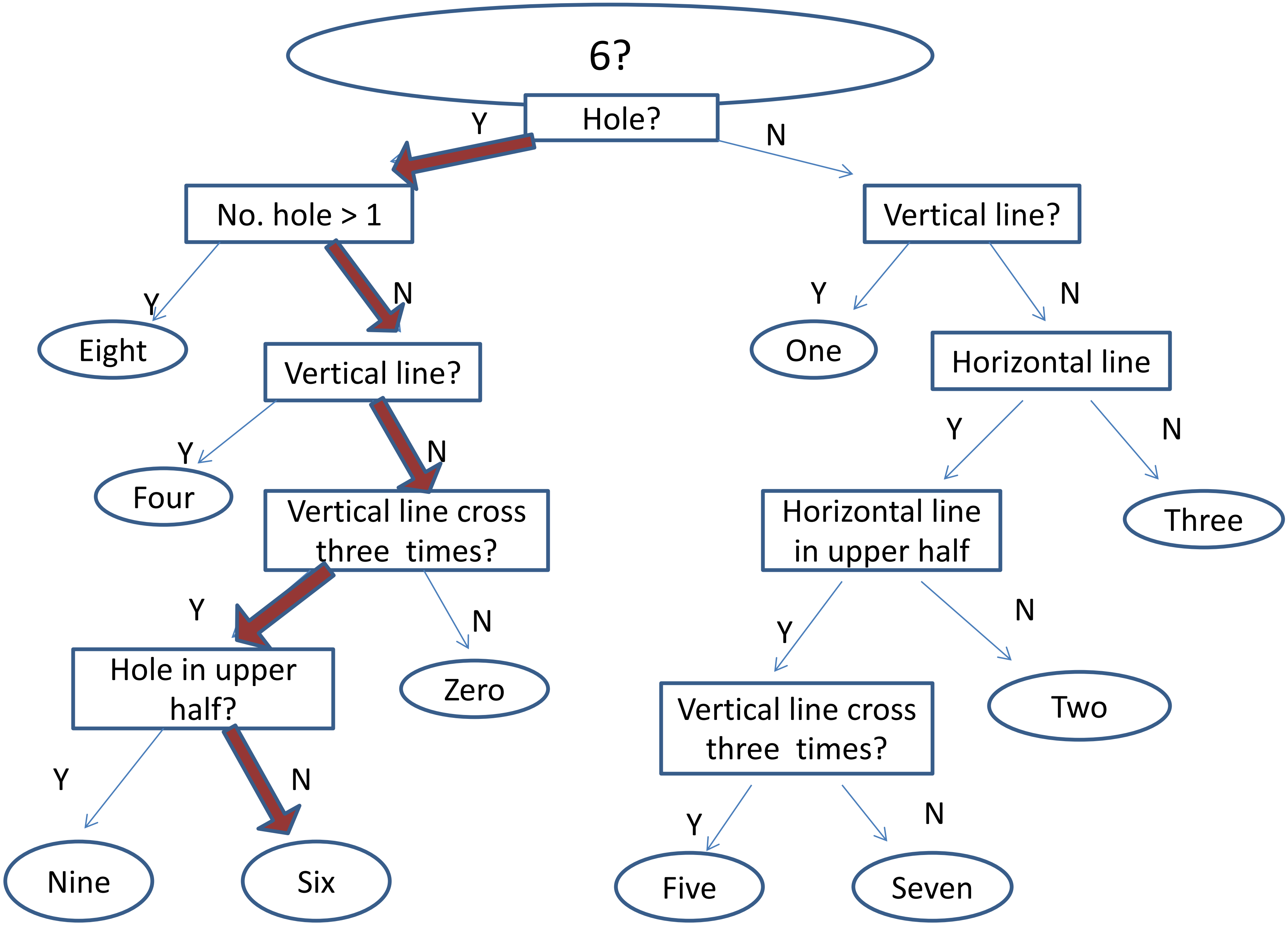
Numerical digit recognition

- Numerical digits: 0 1 2 3 4 5 6 7 8 9




0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Hole?
Y
N
0, 4, 6, 8, 9
No. hole > 1
Y
N
8
0, 4, 6, 9
Vertical line?
Y
N
4
0, 6, 9
Vertical line cross three times?
Y
N
6, 9
0
Hole in upper half?
Y
N
9
6
1, 2, 3, 5, 7,
Vertical line?
Y
N
1
2, 3, 5, 7
Horizontal line
Y
N
2, 5, 7
3
Horizontal line in upper half
Y
N
5, 7
2
Vertical line cross three times?
Y
N
5
7


Hole?
Y
N
No. hole > 1
Y
Eight
N
Vertical line?
Y
Four
N
Vertical line cross three times?
Y
Hole in upper half?
Y
Nine
N
Six
N
Zero
Vertical line?
Y
One
N
Horizontal line
Y
Horizontal line in upper half
Y
Vertical line cross three times?
Y
Five
N
Seven
N
Two
N
Three


6?
Hole?
Y
N
No. hole > 1
Vertical line?
Y
N
Eight
Vertical line?
One
Horizontal line
Y
N
Four
Vertical line cross three times?
Horizontal line in upper half
Three
Y
N
Hole in upper half?
Zero
Vertical line cross three times?
Two
Nine
Six
Five
Seven

- This is called **Tree classifier** or **Decision tree** classifier.
- At each node you have to take a decision.
- Through some sequence of decisions we can build a classifier.
- Leaf node gives the label.

Why tree

- Simplicity of design
- Interpretability
- Ease of implementation
- Good performance in practice

There are several questions?

- How many decision outcomes or splits will be at node?
- Which feature should be tested at a node?
- When should a node be declare a leaf node?
- If the tree becomes too large, how can it be made smaller and simpler?
- If a leaf node is impure, how should the category label be assigned?
- How should missing data be handled?

How many decision outcomes or splits?

- 1, 2, 3,……., C
- Every decision can be represented using just binary decision.
- Binary tree (widely used)

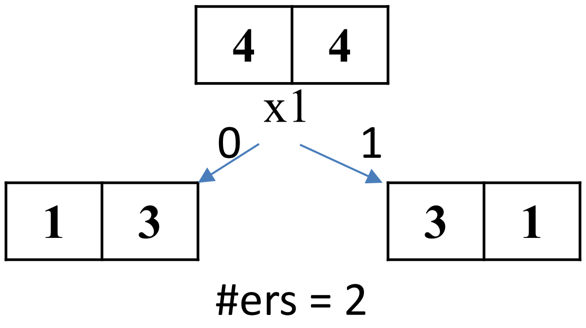
  - easy to handle

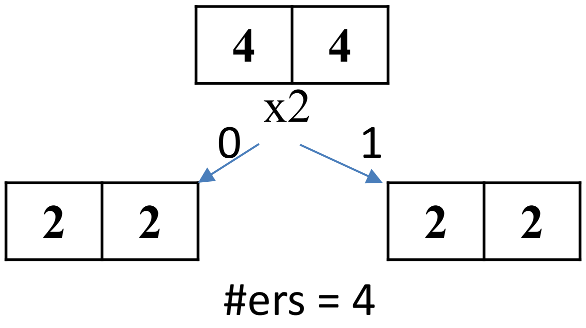
  - theoretical development

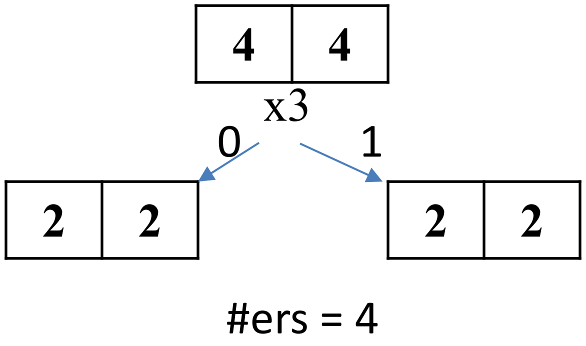
Which feature should be tested at a node?

- Much of the work on decision tree to answer this question

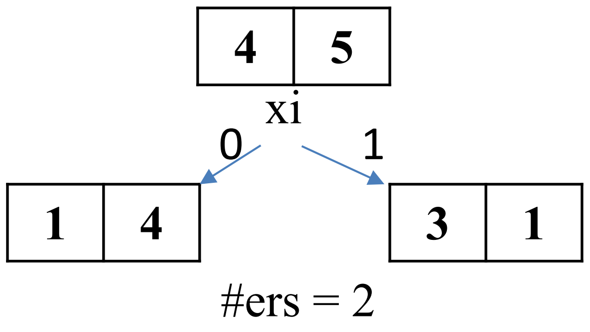
x1	x2	x3	y
0	0	0	1
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	0

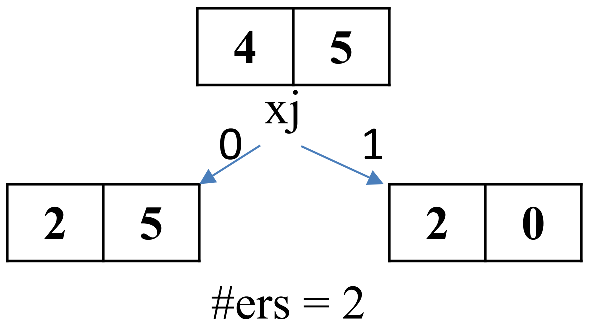






Which feature should be tested at a node?





- Which one (xi or xj) should be tested?
- Why?

Better Heuristics

- Let X be a random variable with the following probability distribution:

P(X=0)	**P(X=1)**
0.1	0.9

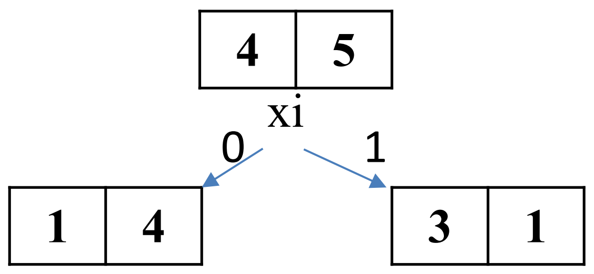
- Define *surprise* of each values of X as:

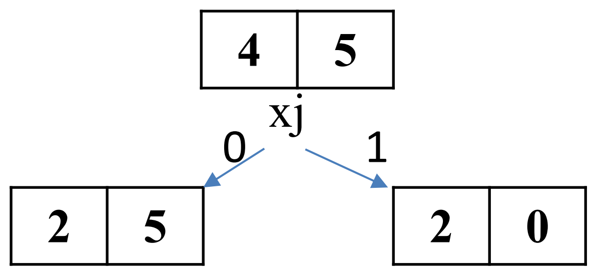
$$S(X=x) = -\log(P(X=x))$$

- ***Entropy*** (*average surprise*):

$$H(X) = -\sum_{x=0}^{1} P(X=x)\log(P(X=x))$$

$$H(X) = -(4/9)\log(4/9) - (5/9)\log(5/9) = 0.9911$$





$$H(L\_xi) = -(1/5)\log(1/5) - (4/5)\log(4/5) = 0.7219$$

$$H(R\_xi) = -(3/4)\log(3/4) - (1/4)\log(1/4) = 0.8113$$

$$H(X|xi) = (5/9)H(L\_Xi) + (4/9)H(R\_xi) = 0.7847$$

$$IG(xi) = H(X) - H(X|xi) = 0.9911-0.7847 = 0.2064$$

$$H(L\_xj) = -(2/7)\log(2/7) - (5/7)\log(5/7) = 0.8631$$

$$H(R\_xj) = -(2/2)\log(2/2) - (0/2)\log(0/2) = 0$$

$$H(X|xj) = (7/9)H(L\_Xj) + (2/9)H(R\_xj) = 0.6713$$

$$IG(xj) = H(X) - H(X|xj) = 0.9911-0.6713 = 0.3198$$

Entropy impurity

- If a node has a proportion of p_j of each of the classes then the information or entropy is:

$$i(N) = -\sum_j p_j \log p_j$$

where 0log0 = 0

Gini impurity

- This is the most widely used measure of impurity
- Gini index is:

$$i(N) = \sum_{i \neq j} p_i p_j = 1 - \sum_j p_j^2$$

Feature selection at a node

- Best query value s is the choice for feature that maximize $\Delta i(N)$

$$\Delta i(N) = i(N) - \sum_{k=1}^{B} P_k i(N_k)$$

Where P_k ‘s are the fraction of patterns at node N that will go to N_k

When should a node be declare a leaf node?

- Impurity is zero
  - over fitting
- Threshold on number of points
  - difficult to choose

If the tree becomes too large, how can it be made smaller and simpler?

- Pruning
- The pruning is based on the misclassification rate. However the error rate will always drop (or at least not increase) with every split. This does not mean that the error rate on test data will improve.
- The solution to this problem is **cross-validation**.

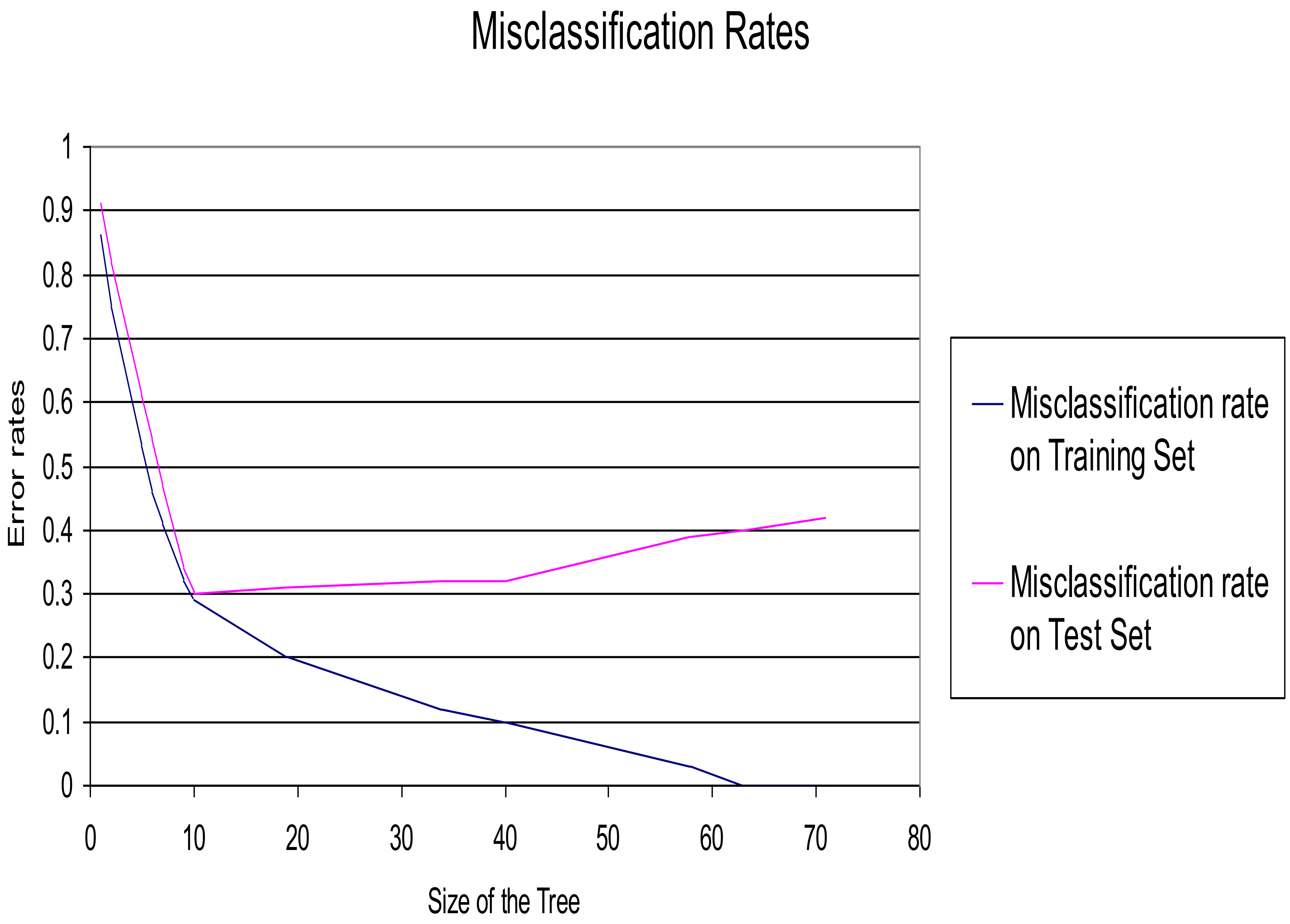
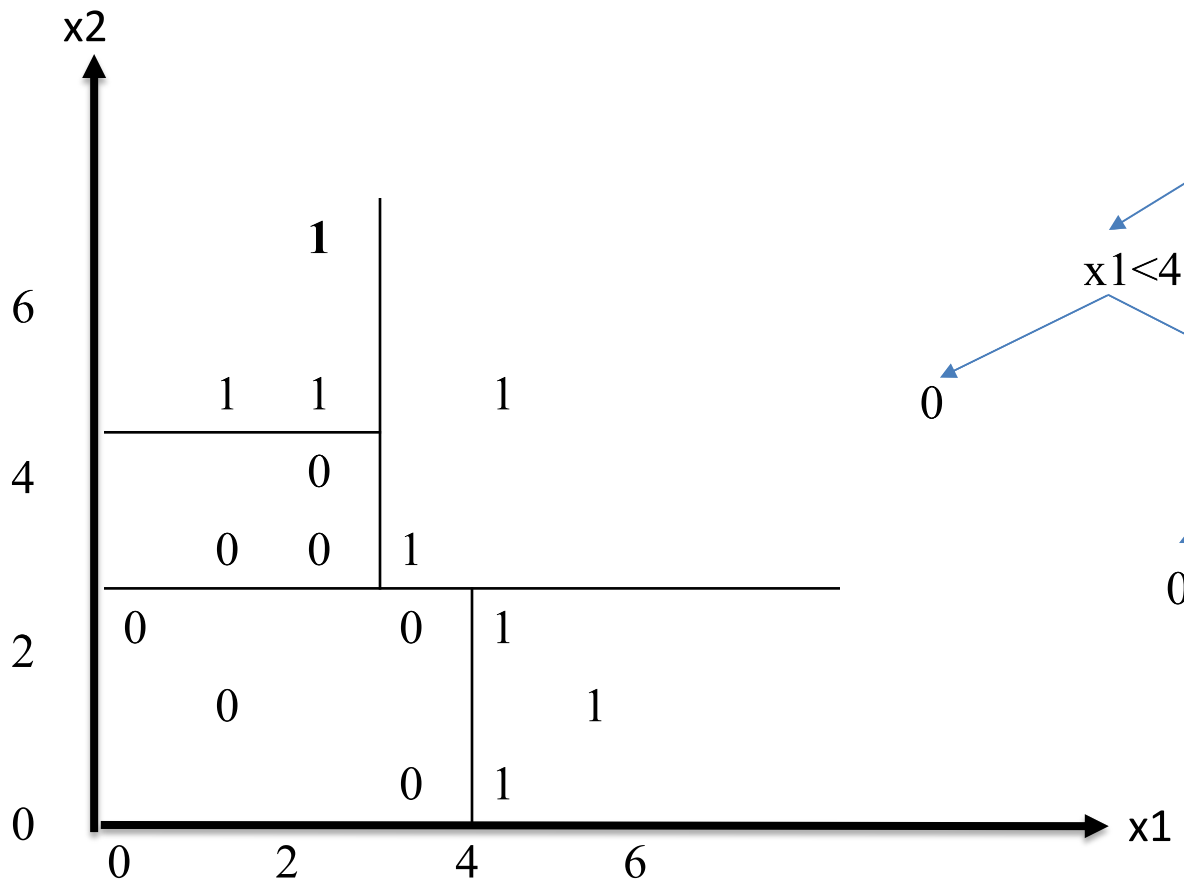
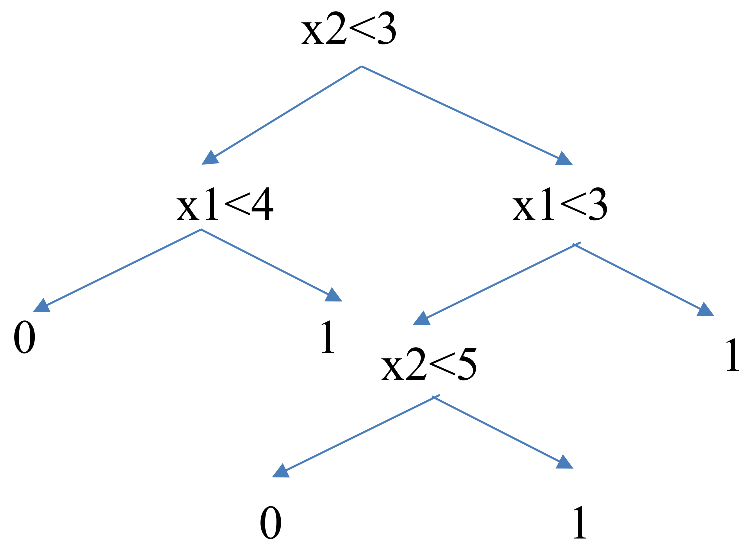


Image: CART by Breiman et al.

- If a leaf node is impure, how should the category label be assigned?
  - majority voting
- How should missing data be handled?
  - delete that feature
  - interpolation

Decision Tree Boundaries





References

- Please follow the suggested reading material