COMP7/8118 M50

Data Mining

Decision Tree: II

Xiaofei Zhang

Slides compiled from Jiawei Han and Raymond C.W. Wong's work

THE UNIVERSITY OF MEMPHIS

Decision Tree Based Classification

- Advantages:
  - Inexpensive to construct
  - Extremely fast at classifying unknown records
  - Easy to interpret for small-sized trees
  - Accuracy is comparable to other classification techniques for many simple data sets

Decision Tree Examples

- ID3
  - Impurity Measurement
    - Gain(A, T) = Info(T) – Info(A, T)
- C4.5
  - Impurity Measurement
    - Gain(A, T) = (Info(T) – Info(A, T))/SplitInfo(A), where SplitInfo(A) = $-\Sigma_{v \in A}$ p(v) log p(v)
- CART
  - Impurity Measurement
    - Gini: I(P) = $1 - \Sigma_j p_j^2$

Iterative Dichotomiser

Classification

Classification And Regression Trees

Entropy

Gender	Income	Child	Insurance
female	high	no	yes
male	high	yes	yes
male	low	yes	yes
male	low	yes	yes
female	low	no	no
female	low	no	no
female	low	no	no
male	low	no	no

$$\text{Info}(T) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} = 1$$

For attribute Gender,

$$\text{Info}(T_{female}) = -\tfrac{3}{4}\log\tfrac{3}{4} - \tfrac{1}{4}\log\tfrac{1}{4} = 0.8113$$

$$\text{Info}(T_{male}) = -\tfrac{3}{4}\log\tfrac{3}{4} - \tfrac{1}{4}\log\tfrac{1}{4} = 0.8113$$

$$\text{Info}(\text{Gender}, T) = \tfrac{1}{2} \times \text{Info}(T_{male}) + \tfrac{1}{2} \times \text{Info}(T_{female}) = 0.8113$$

$$\text{Gain}(\text{Gender}, T) = \text{Info}(T) - \text{Info}(\text{Gender}, T) = 1 - 0.8113 = 0.1887$$

For attribute Gender, Gain(Gender, T) = 0.1887

Entropy

Gender	Income	Child	Insurance
female	high	no	yes
male	high	yes	yes
male	low	yes	yes
male	low	yes	yes
female	low	no	no
female	low	no	no
female	low	no	no
male	low	no	no

$$\text{Info}(T) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} = 1$$

For attribute Income,

$$\text{Info}(T_{high}) = -1\log 1 - 0\log 0 = 0$$

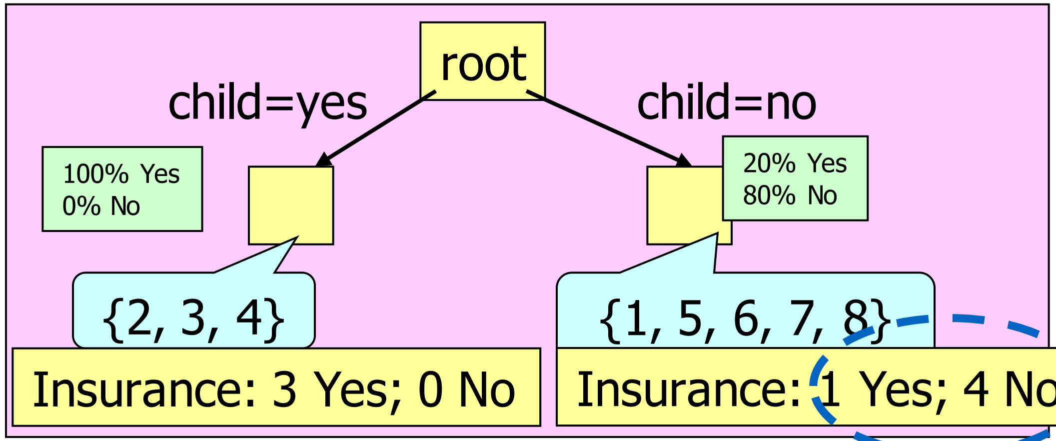
$$\text{Info}(T_{low}) = -1/3\log 1/3 - 2/3\log 2/3 = 0.9183$$

$$\text{Info}(\text{Income}, T) = \tfrac{1}{4} \times \text{Info}(T_{high}) + \tfrac{3}{4} \times \text{Info}(T_{low}) = 0.6887$$

$$\text{Gain}(\text{Income}, T) = \text{Info}(T) - \text{Info}(\text{Income}, T) = 1 - 0.6887 = 0.3113$$

For attribute Gender, Gain(Gender, T) = 0.1887

For attribute Income, Gain(Income, T) = 0.3113



	Gender	Income	Child	Insurance
1	female	high	no	yes
2	male	high	yes	yes
3	male	low	yes	yes
4	male	low	yes	yes
5	female	low	no	no
6	female	low	no	no
7	female	low	no	no
8	male	low	no	no

$$\text{Info}(T) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} = 1$$

For attribute Child,

$$\text{Info}(T_{yes}) = -1\log 1 - 0\log 0 = 0$$

$$\text{Info}(T_{no}) = -1/5\log 1/5 - 4/5\log 4/5 = 0.7219$$

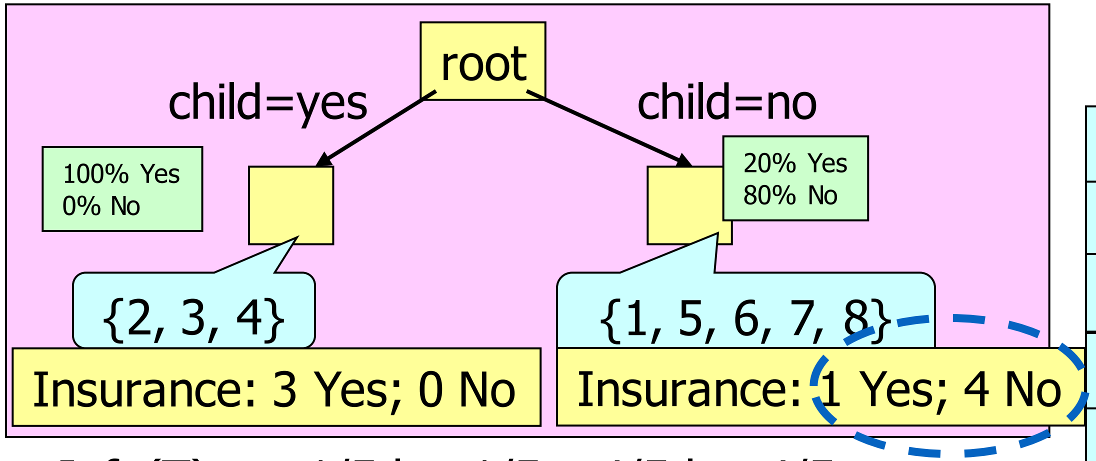
$$\text{Info}(\text{Child}, T) = 3/8 \times \text{Info}(T_{yes}) + 5/8 \times \text{Info}(T_{no}) = 0.4512$$

$$\text{Gain}(\text{Child}, T) = \text{Info}(T) - \text{Info}(\text{Child}, T) = 1 - 0.4512 = 0.5488$$

For attribute Gender, Gain(Gender, T) = 0.1887

For attribute Income, Gain(Income, T) = 0.3113

For attribute Child, Gain(Child, T) = 0.5488



	Gender	Income	Child	Insurance
1	female	high	no	yes
2	male	high	yes	yes
3	male	low	yes	yes
4	male	low	yes	yes
5	female	low	no	no
6	female	low	no	no
7	female	low	no	no
8	male	low	no	no

$$\text{Info}(T) = -\ 1/5 \log 1/5 - 4/5 \log 4/5$$
$$= 0.7219$$

For attribute Gender,

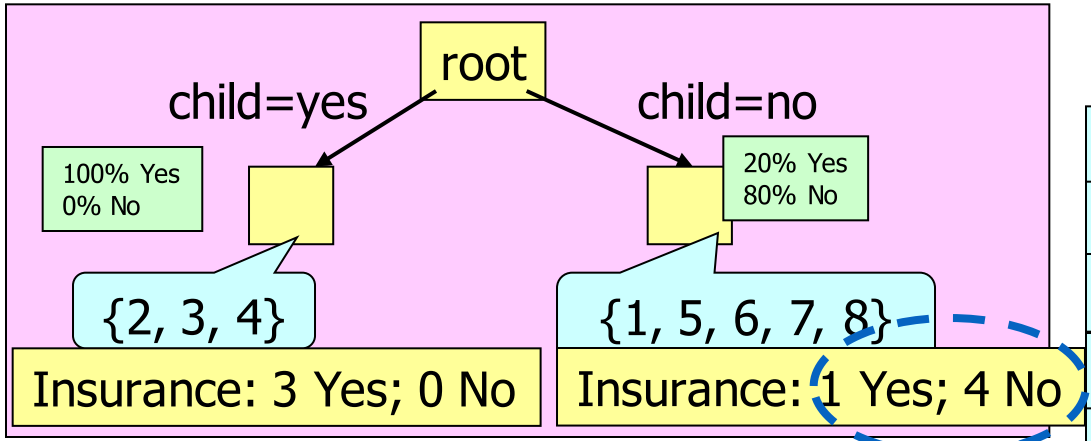
$$\text{Info}(T_{female}) = -\ ¼ \log ¼ - ¾ \log ¾ = 0.8113$$

$$\text{Info}(T_{male}) = -\ 0 \log 0 - 1 \log 1 = 0$$

$$\text{Info}(\text{Gender}, T) = 4/5 \times \text{Info}(T_{female}) + 1/5 \times \text{Info}(T_{male}) = 0.6490$$

$$\text{Gain}(\text{Gender}, T) = \text{Info}(T) - \text{Info}(\text{Gender}, T) = 0.7219 - 0.6490 = 0.0729$$

For attribute Gender, Gain(Gender, T) = 0.0729



	Gender	Income	Child	Insurance
1	female	high	no	yes
2	male	high	yes	yes
3	male	low	yes	yes
4	male	low	yes	yes
5	female	low	no	no
6	female	low	no	no
7	female	low	no	no
8	male	low	no	no

Info(T) = - 1/5 log 1/5 – 4/5 log 4/5
= 0.7219

For attribute Income,

$\text{Info}(T_{high}) = - 1 \log 1 - 0 \log 0 = 0$

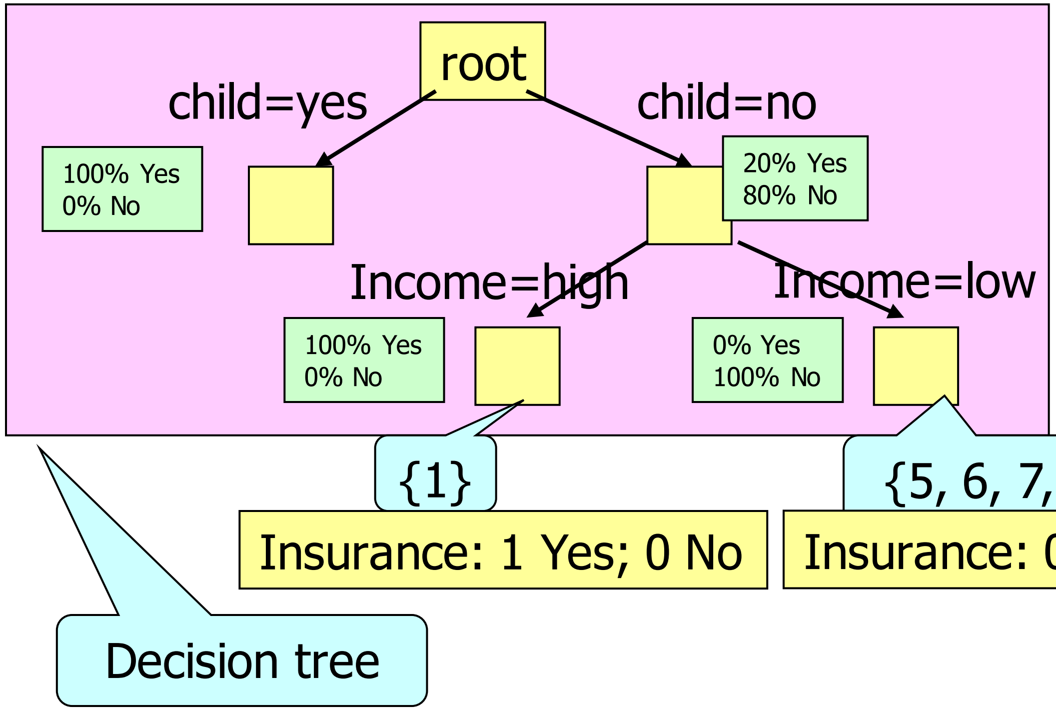
$\text{Info}(T_{low}) = - 0 \log 0 - 1 \log 1 = 0$

$\text{Info}(\text{Income}, T) = 1/5 \times \text{Info}(T_{high}) + 4/5 \times \text{Info}(T_{low}) = 0$

Gain(Income, T) = Info(T) – Info(Income, T) = 0.7219 – 0 = 0.7219

For attribute Gender, Gain(Gender, T) = 0.0729

For attribute Income, Gain(Income, T) = 0.7219



	Gender	Income	Child	Insurance
1	female	high	no	yes
2	male	high	yes	yes
3	male	low	yes	yes
4	male	low	yes	yes
[illegible]	female	low	no	no
[illegible]	[illegible]	low	no	no
[illegible]	female	low	no	no
8	male	low	no	no

Suppose there is a new person.

Gender	Income	Child	Insurance
male	high	no	?

Decision Tree Examples

- ID3
  - Impurity Measurement
    - Gain(A, T) = Info(T) – Info(A, T)
- C4.5
  - Impurity Measurement
    - Gain(A, T) = (Info(T) – Info(A, T))/SplitInfo(A), where SplitInfo(A) = $-\Sigma_{v \in A}$ p(v) log p(v)
- CART
  - Impurity Measurement
    - Gini: $I(P) = 1 - \Sigma_j p_j^2$

Entropy

Gender	Income	Child	Insurance
female	high	no	yes
male	high	yes	yes
male	low	yes	yes
male	low	yes	yes
female	low	no	no
female	low	no	no
female	low	no	no
male	low	no	no

$\text{Info}(T) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$
$= 1$

For attribute Gender,

$\text{Info}(T_{female}) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113$

$\text{Info}(T_{male}) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113$

$\text{Info}(\text{Gender}, T) = \frac{1}{2} \times \text{Info}(T_{female}) + \frac{1}{2} \times \text{Info}(T_{male}) = 0.8113$

$\text{SplitInfo}(\text{Gender}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$\text{Gain}(\text{Gender}, T) = (\text{Info}(T) - \text{Info}(\text{Gender}, T))/\text{SplitInfo}(\text{Gender}) = (1 - 0.8113)/1 = 0.1887$

For attribute Gender, Gain(Gender, T) = 0.1887

Entropy

Gender	Income	Child	Insurance
female	high	no	yes
male	high	yes	yes
male	low	yes	yes
male	low	yes	yes
female	low	no	no
female	low	no	no
female	low	no	no
male	low	no	no

$\text{Info}(T) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$
$= 1$

For attribute Income,

$\text{Info}(T_{high}) = -1\log 1 - 0\log 0 = 0$

$\text{Info}(T_{low}) = -1/3\log 1/3 - 2/3\log 2/3 = 0.9183$

$\text{Info}(\text{Income}, T) = \frac{1}{4} \times \text{Info}(T_{high}) + \frac{3}{4} \times \text{Info}(T_{low}) = 0.6887$

$\text{SplitInfo}(\text{Income}) = -2/8\log 2/8 - 6/8\log 6/8 = 0.8113$

$\text{Gain}(\text{Income}, T) = (\text{Info}(T) - \text{Info}(\text{Income}, T))/\text{SplitInfo}(\text{Income}) = (1-0.6887)/0.8113$
$= 0.3837$

For attribute Gender, $\text{Gain}(\text{Gender}, T) = 0.1887$

For attribute Income, $\text{Gain}(\text{Income}, T) = 0.3837$

For attribute Child, $\text{Gain}(\text{Child}, T) = ?$

Decision Tree Examples

- ID3
  - Impurity Measurement
    - Gain(A, T) = Info(T) – Info(A, T)
- C4.5
  - Impurity Measurement
    - Gain(A, T) = (Info(T) – Info(A, T))/SplitInfo(A), where SplitInfo(A) = $-\Sigma_{v \in A}$ p(v) log p(v)
- CART
  - Impurity Measurement
    - Gini: I(P) = $1 - \Sigma_j p_j^2$

Gini

Gender	Income	Child	Insurance
female	high	no	yes
male	high	yes	yes
male	low	yes	yes
male	low	yes	yes
female	low	no	no
female	low	no	no
female	low	no	no
male	low	no	no

$$\text{Info}(T) = 1 - (½)^2 - (½)^2$$
$$= ½$$

For attribute Gender,

$$\text{Info}(T_{female}) = 1 - (¾)^2 - (¼)^2 = 0.375$$

$$\text{Info}(T_{male}) = 1 - (¾)^2 - (¼)^2 = 0.375$$

$$\text{Info}(\text{Gender}, T) = ½ \text{ x Info}(T_{female}) + ½ \text{ x Info}(T_{male}) = 0.375$$

$$\text{Gain}(\text{Gender}, T) = \text{Info}(T) - \text{Info}(\text{Gender}, T) = ½ - 0.375 = 0.125$$

For attribute Gender, Gain(Gender, T) = 0.125

Gini

Gender	Income	Child	Insurance
female	high	no	yes
male	high	yes	yes
male	low	yes	yes
male	low	yes	yes
female	low	no	no
female	low	no	no
female	low	no	no
male	low	no	no

$\text{Info}(T) = 1 - (½)^2 - (½)^2$
$= ½$

For attribute Income,

$\text{Info}(T_{high}) = 1 - 1^2 - 0^2 = 0$

$\text{Info}(T_{low}) = 1 - (1/3)^2 - (2/3)^2 = 0.444$

$\text{Info}(\text{Income}, T) = 1/4 \times \text{Info}(T_{high}) + 3/4 \times \text{Info}(T_{low}) = 0.333$

$\text{Gain}(\text{Income}, T) = \text{Info}(T) - \text{Info}(\text{Gender}, T) = ½ - 0.333 = 0.167$

For attribute Gender, Gain(Gender, T) = 0.125

For attribute Income, Gain(Gender, T) = 0.167

For attribute Child, Gain(Child, T) = ?

Other Issues

- Data Fragmentation
- Expressiveness

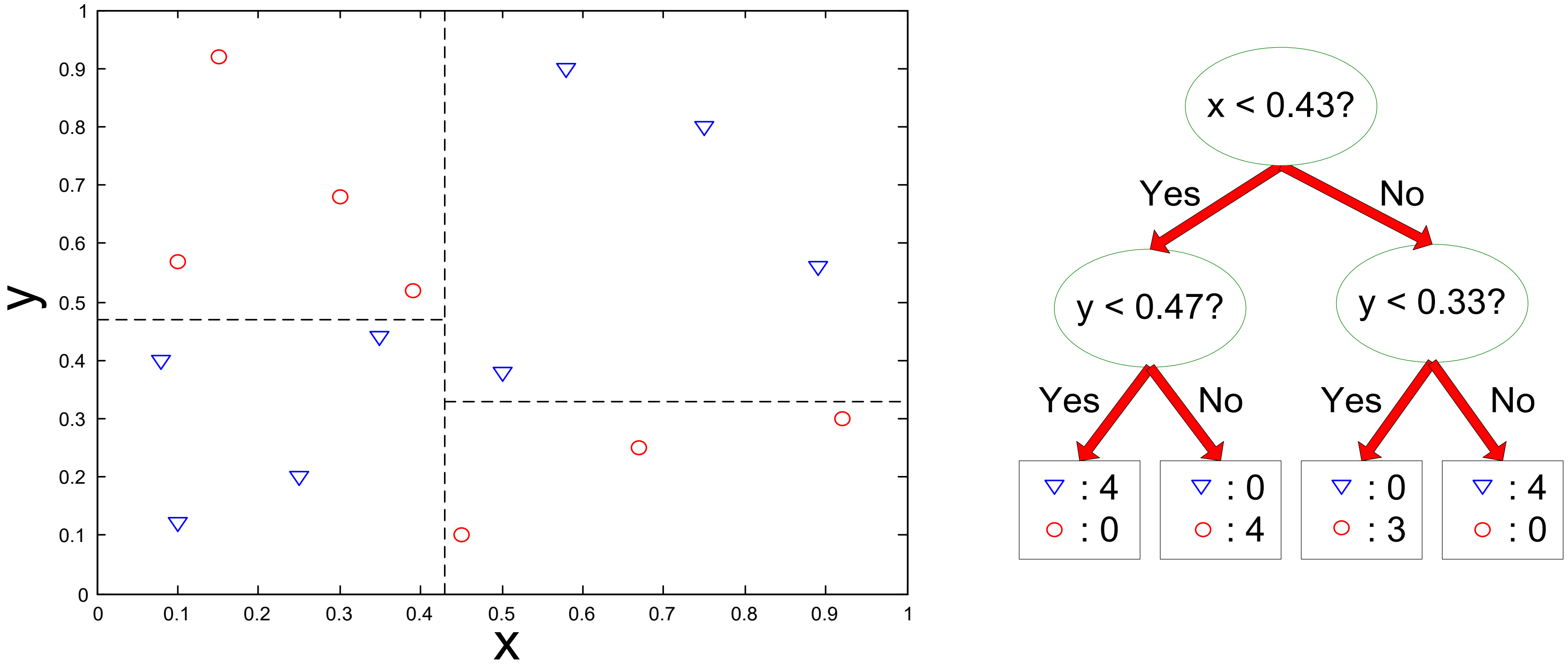
Data Fragmentation

- Number of instances gets smaller as you traverse down the tree
- Number of instances at the leaf nodes could be too small to make any statistically significant decision
- You can introduce a lower bound on the number of items per leaf node in the stopping criterion.

Expressiveness

- A classifier defines a function that discriminates between two (or more) classes.
- The expressiveness of a classifier is the class of functions that it can model, and the kind of data that it can separate
    - When we have discrete (or binary) values, we are interested in the class of boolean functions that can be modeled
    - If the data-points are real vectors we talk about the decision boundary that the classifier can model

Decision Boundary



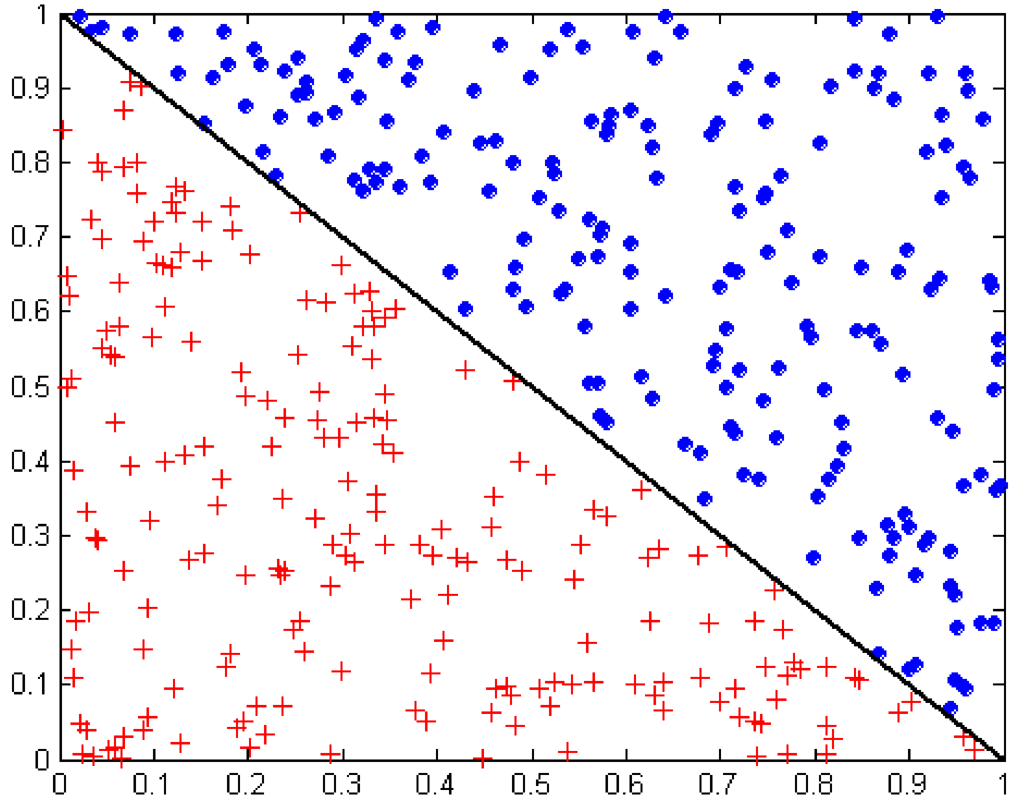
- Border line between two neighboring regions of different classes is known as decision boundary
- Decision boundary is parallel to axes because test condition involves a single attribute at-a-time

Expressiveness

- Decision tree provides expressive representation for learning discrete-valued function
  - But they do not generalize well to certain types of Boolean functions
    - Example: parity function:
      - Class = 1 if there is an even number of Boolean attributes with truth value = True
      - Class = 0 if there is an odd number of Boolean attributes with truth value = True
    - For accurate modeling, must have a complete tree

- Less expressive for modeling continuous variables
  - Particularly when test condition involves only a single attribute at-a-time

Oblique Decision Trees



- Test condition may involve multiple attributes
- More expressive representation
- Finding optimal test condition is computationally expensive