香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

COMP7015 Artificial Intelligence

Lecture 5: Machine Learning I --- Linear Models and Decision Tree

Instructor: Dr. Kejing Yin

October 6, 2022

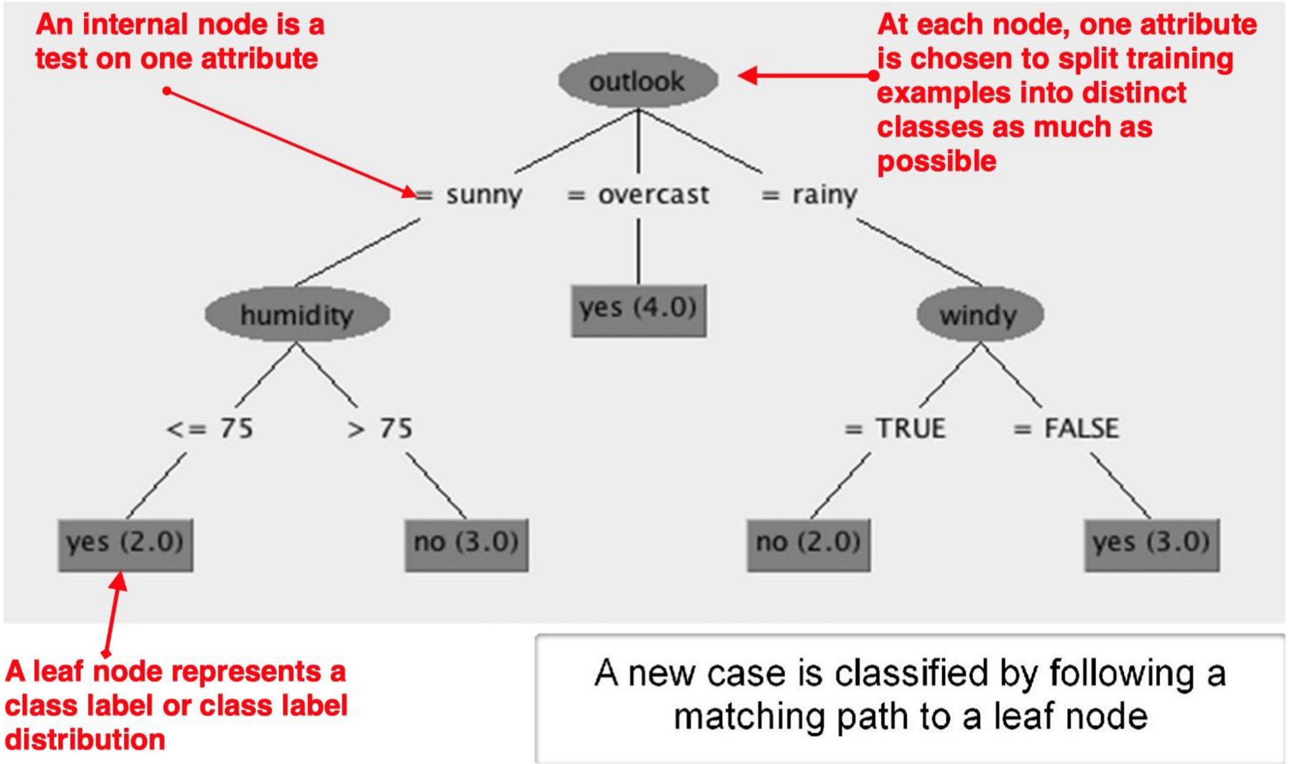
Decision Tree Algorithm

Play Tennis or Not?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Given past data:

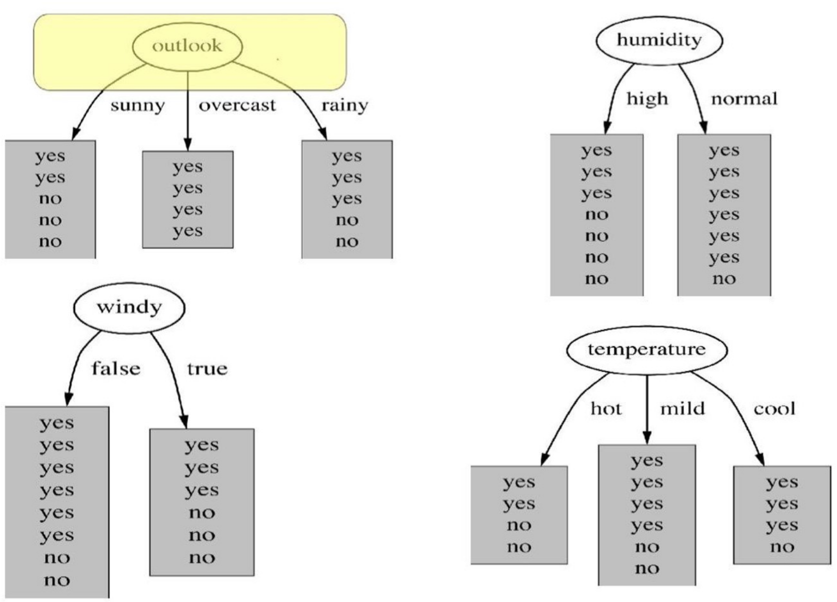
build a system to predict **Play/Not Play** ?



- Decision Tree is a method for approximating classification functions by means of a **tree-based representation**.
- A learned Decision Tree can be represented as a set of **if-then** rules

Idea: building a classification tree

- **Top-down tree construction:**
  - At start, all training samples are at the root.
  - Split the samples recursively by choosing one **attribute** each time

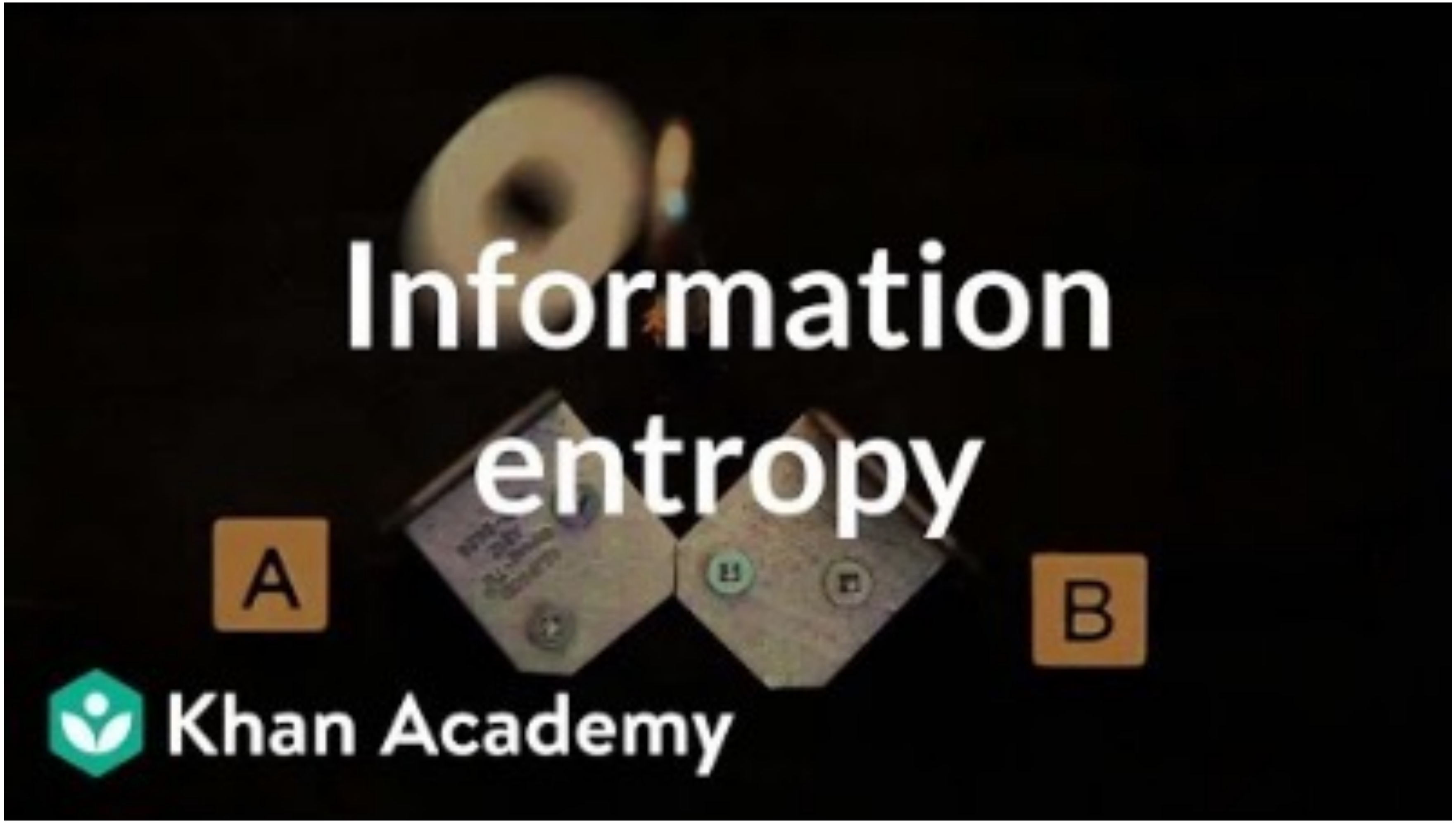


Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

  - Which attribute should we choose at each node?
  - Divide-and-conquer: Split the into smaller subset.
  - criteria: choose the attribute that **"best" separates** the classes on the training samples
  - How do we measure the information contained in each attribute? information measure

Information measures: Entropy

- Entropy is a measure of the uncertainty of a random variable.



https://www.youtube.com/watch?v=2s3aJfRr9gE

Information measures: Entropy

- Entropy is a measure of the uncertainty of a random variable (or purity of a dataset).

- Given dataset D with K classes of samples, e.g., play/not play data: $K = 2$

- Entropy of dataset D: $$\boxed{\mathrm{Ent}(D) = -\sum_{k=1}^{K} p_k \log_2 p_k}$$

  p_k is the frequency of the k-th class

- The smaller $\mathrm{Ent}(D)$, the purer D.

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Exercise: compute $\mathrm{Ent}(D)$

$\log_2 0.643 \approx -0.637$ $\log_2 0.357 \approx -1.486$

- $K = 2$ (Yes/No)
- Samples with $k = 1$ (Yes):
  $p_k = 9/14 = 0.643$
- Samples with $k = 2$ (No):
  $p_k = 5/14 = 0.357$

$$\begin{aligned}\mathrm{Ent}(D) &= -(0.643 * \log_2 0.643 + 0.357 * \log_2 0.357) \\ &= -(0.643 * -0.637 + 0.357 * -1.486) \\ &= 0.94\end{aligned}$$

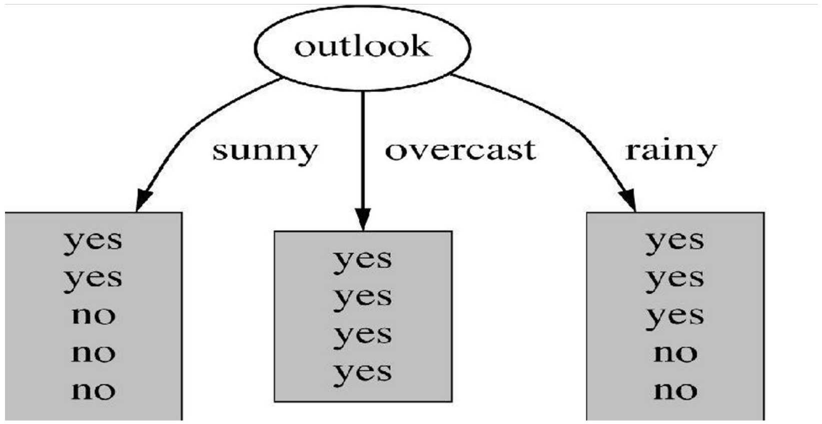
Exercise: compute $\mathrm{Ent}(D)$ for the following datasets

(only labels are shown)

- $D_1 = \{+1, +1, -1, -1, -1, +1, -1, +1\}$
- $D_2 = \{A, B, B, A, C, D, B, A, C, D, D, C\}$
- $D_3 = \{A, A, A, A, A, B, A, A, C, C, D, A\}$

$$\log_2 0.5 = -1$$
$$\log_2 0.25 = -2$$
$$\log_2 0.67 \approx -0.578$$
$$\log_2 0.083 \approx -3.591$$
$$\log_2 0.167 \approx -2.582$$

Information Gain



- For attribute a, it has V possible values.
  E.g., a="outlook", $V = 3$

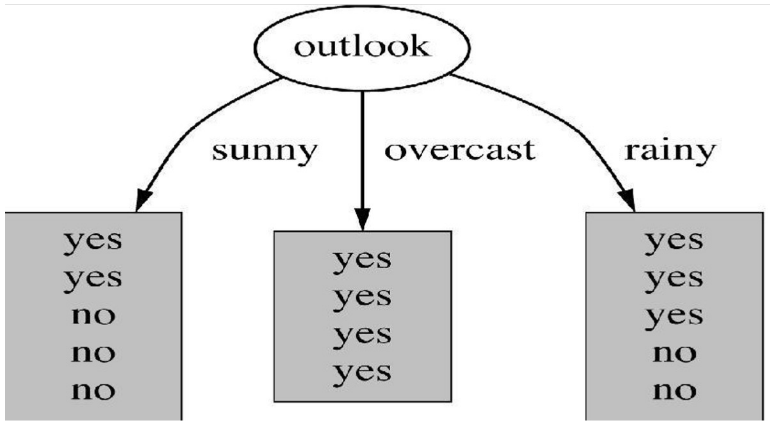
- If we divide the data using a, the information gain is:

$$\mathrm{Gain}(D, a) = \boxed{\mathrm{Ent}(D)} - \boxed{\sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)}$$

D^v: dataset that has value of v in D

purity before split *purity after split*

Information Gain



- For attribute a, it has V possible values.
  E.g., a="outlook", $V = 3$

- If we divide the data using a, the information gain is:

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

D^v: dataset that has value of v in D

- What is the information gain $\mathrm{Gain}(D, \text{"outlook"})$?

Sunny: $\frac{|D^v|}{D} \mathrm{Ent}(D^v) = \frac{5}{14}\left(-\frac{2}{5} * \log_2 \frac{2}{5} - \frac{3}{5} * \log_2 \frac{3}{5}\right) = 0.347$

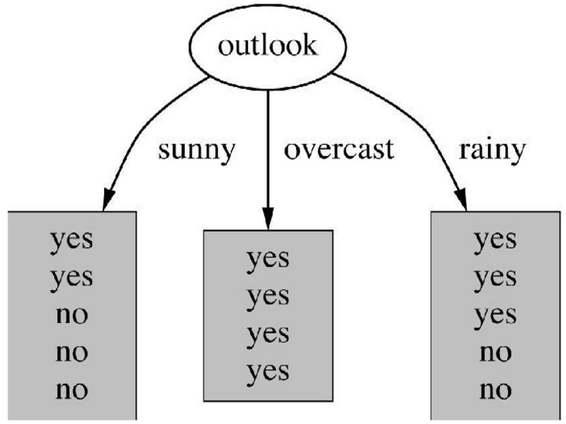
Overcast: $\frac{|D^v|}{D} \mathrm{Ent}(D^v) = \frac{4}{14}\left(-\frac{4}{4} * \log_2 \frac{4}{4} - \frac{0}{4} * \log_2 \frac{0}{4}\right) = 0$ *We define $0\, log_2\, 0 = 0$*

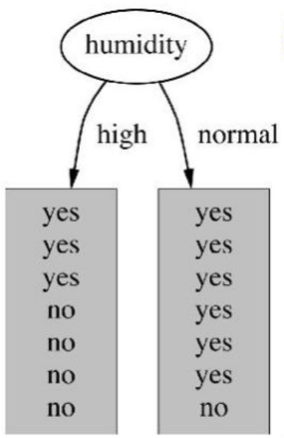
Rainy: $\frac{|D^v|}{D} \mathrm{Ent}(D^v) = \frac{5}{14}\left(-\frac{3}{5} * \log_2 \frac{3}{5} - \frac{2}{5} * \log_2 \frac{2}{5}\right) = 0.347$

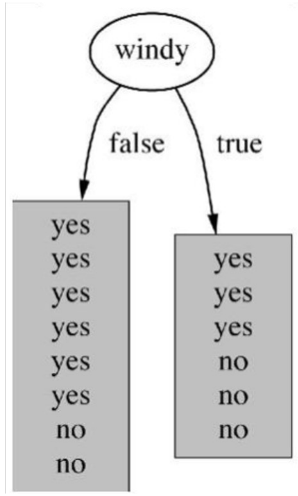
Information Gain: $\mathrm{Gain}(D, \text{"outlook"}) = 0.94 - 0.347 - 0 - 0.347 = 0.246$

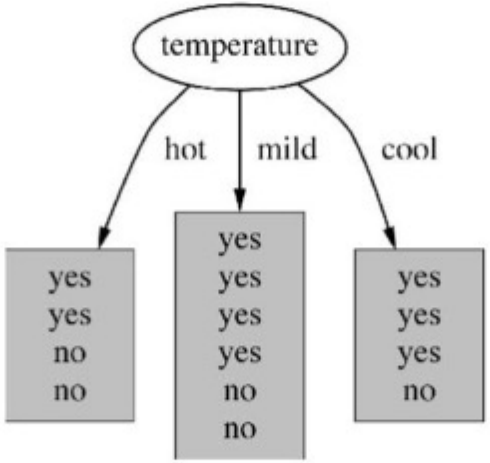
Compute the information gain for other attributes

highest









$\text{Gain}(D, \text{“outlook”}) = 0.246$

$\text{Gain}(D, \text{“windy”}) = 0.048$

$\text{Gain}(D, \text{“humidity”}) = 0.152$

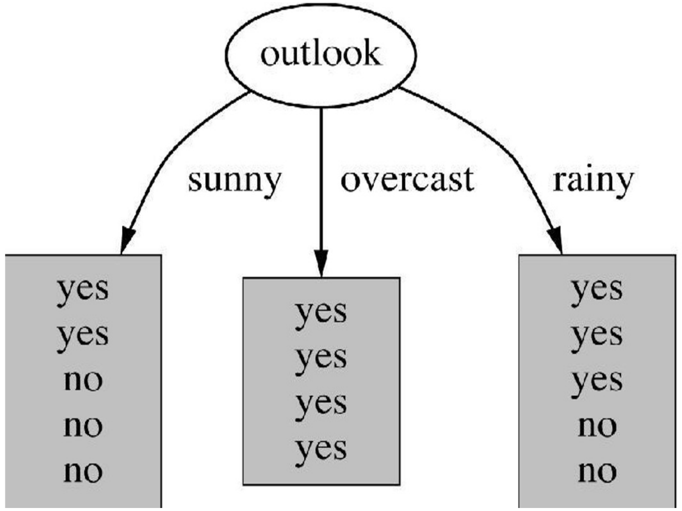
$\text{Gain}(D, \text{“temperature”}) = 0.029$

We can split data using the attribute that has the highest information gain

Split by information gain: ID3 algorithm

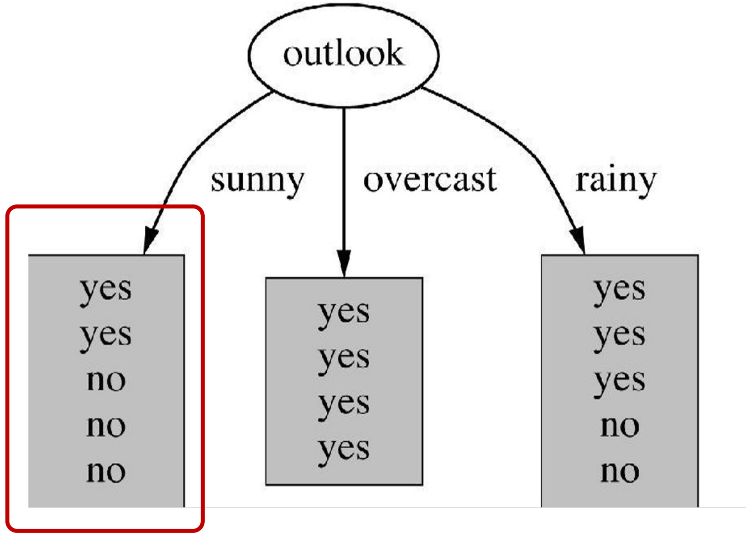
```
ID3(D,X) =
  Let T be a new tree
  If all instances in D have same class c
    Label(T) = c; Return T
  If X = ∅ or no attribute has positive information gain
    Label(T) = most common class in D; return T
  X ← attribute with highest information gain
  Label(T) = X
  For each value x of X
    D_x ← instances in D with X = x
    If D_x is empty
      Let T_x be a new tree
      Label(T_x) = most common class in D
    Else
      T_x = ID3(D_x, X – { X })
    Add a branch from T to T_x labeled by x
  Return T
```

Recursively call the ID3 algorithm (as if this is a brand new dataset)



Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

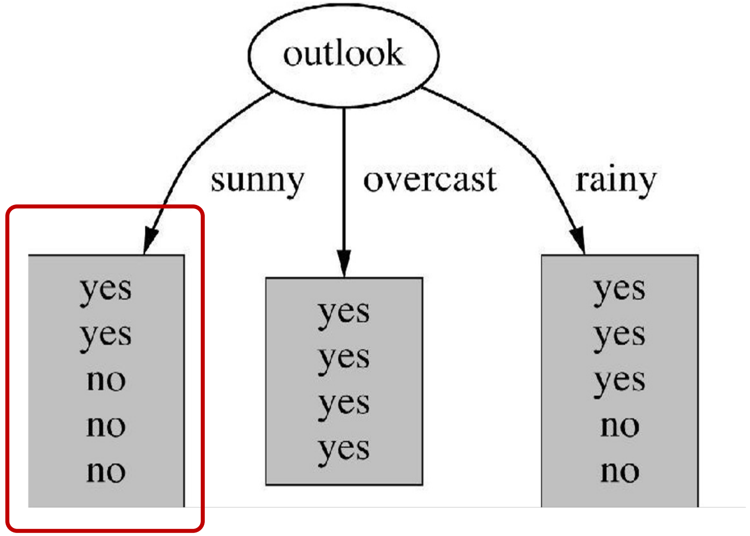
Step 2



- When outlook=sunny

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
sunny	mild	normal	true	Yes

Step 2



- As if we are given this dataset and continue selecting the attribute to split this "new" dataset.

$$\mathrm{Ent}(D) = -(\frac{3}{5} * \log_2 \frac{3}{5} + \frac{2}{5} * \log_2 \frac{2}{5}) = 0.971$$

- Information Gain of the remaining attributes (temperature, humidity, windy)?

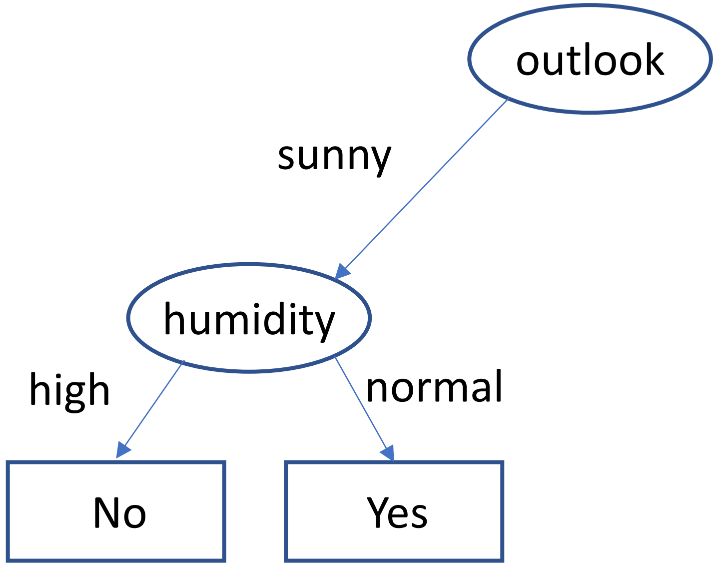
$$\mathrm{Gain}(D, \text{“temperature”}) = 0.971 - \frac{2}{5}\mathrm{Ent}\left(D^{t=\mathrm{hot}}\right) - \frac{2}{5}\mathrm{Ent}\left(D^{\mathrm{t=mild}}\right) - \frac{1}{5}\mathrm{Ent}\left(D^{\mathrm{t=cool}}\right) = 0.971 - 0 - \frac{2}{5} - 0 = 0.571$$

$$\boxed{\mathrm{Gain}(D, \text{“humidity”}) = 0.971 - \frac{3}{5}\mathrm{Ent}\left(D^{\mathrm{h=high}}\right) - \frac{2}{5}\mathrm{Ent}\left(D^{\mathrm{h=normal}}\right) = 0.971 - 0 - 0 = 0.971}$$

$$\mathrm{Gain}(D, \text{“windy”}) = 0.971 - \frac{3}{5}\mathrm{Ent}\left(D^{\mathrm{w=false}}\right) - \frac{2}{5}\mathrm{Ent}\left(D^{\mathrm{w=true}}\right) = 0.971 - 0.5510 - 0.4 = 0.02$$

Split by information gain: ID3 algorithm

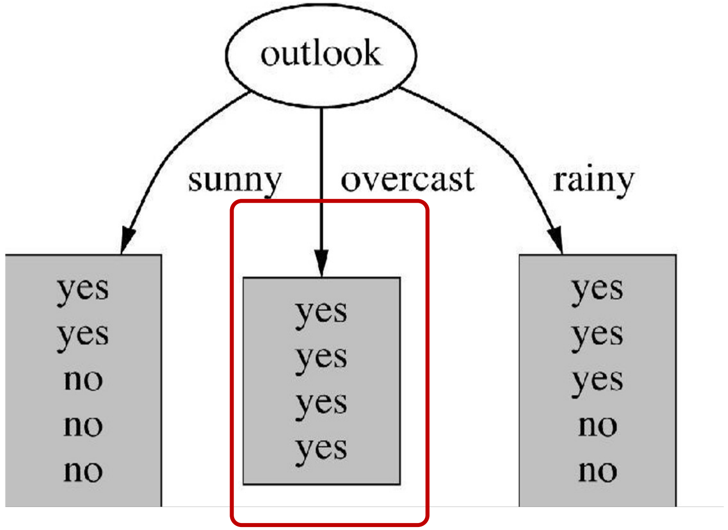
ID3(**D**,**X**) =
    Let T be a new tree
    If all instances in **D** have same class c
        Label(T) = c; Return T
    If **X** = ∅ or no attribute has positive information gain
        Label(T) = most common class in **D**; return T
    $X \leftarrow$ attribute with highest information gain
    Label(T) = X
    For each value x of X
        $\mathbf{D}_x \leftarrow$ instances in **D** with $X = x$
        If $\mathbf{D}_x$ is empty
            Let T_x be a new tree
            Label(T_x) = most common class in **D**
        Else
            $T_x = \mathrm{ID3}(\mathbf{D}_x, \mathbf{X} - \{X\})$
        Add a branch from T to T_x labeled by x
    Return T



Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
sunny	mild	normal	true	Yes

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

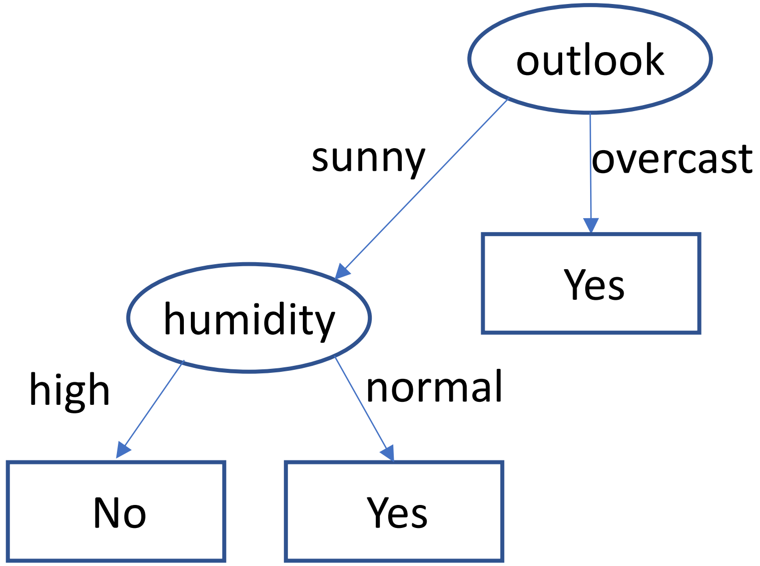
Step 3



- When outlook=overcast

Split by information gain: ID3 algorithm

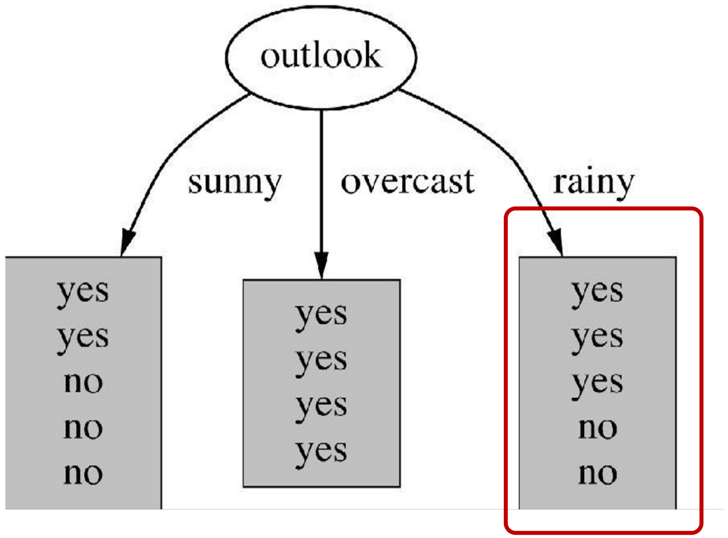
```
ID3(D,X) =
    Let T be a new tree
    If all instances in D have same class c
        Label(T) = c; Return T
    If X = ∅ or no attribute has positive information gain
        Label(T) = most common class in D; return T
    X ← attribute with highest information gain
    Label(T) = X
    For each value x of X
        D_x ← instances in D with X = x
        If D_x is empty
            Let T_x be a new tree
            Label(T_x) = most common class in D
        Else
            T_x = ID3(D_x, X − { X })
        Add a branch from T to T_x labeled by x
    Return T
```



Outlook	Temperature	Humidity	Windy	Play?
overcast	hot	high	false	Yes
overcast	cool	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Step 4



- When outlook=rainy
- Exercise: construct the remaining decision tree.