CS 4104
APPLIED MACHINE LEARNING

Dr. Hashim Yasin

National University of Computer and Emerging Sciences,

Faisalabad, Pakistan.

DECISION TREE (ID3)

Entropy

- **Entropy** characterizes the (im)purity of an arbitrary collection of examples S.

# of possible values of X

$$Entropy(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

Entropy

Example

- Given a **collection S**, containing positive and negative examples of some target concept, the entropy of S relative to this **Boolean classification** is

$$\boxed{\text{Entropy}\ (S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}}$$

- $p_{\oplus}$ is the proportion of positive example in S
- $p_{\ominus}$ is the proportion of negative example in S

Information Gain

- Information Gain measure the effectiveness of an attribute
- It is simply the expected reduction in entropy

$$\text{Gain}(S, A) \equiv \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Where:

- ***Values(A)*** is the set of all possible values for attribute **A**
- $\boldsymbol{S_v}$ is the subset of **S** for which attribute **A** has value **v**.

DECISION TREE (CART)

CART

- Classification And Regression Trees
  - Non-parametric (independent of the statistical distribution of the training data)
  - Can use continuous and non-continuous *predictor* variables
  - Can model continuous (regression trees) or categorical (classification trees) *target* variables
  - computationally rapid and can provide high quality classification results

CART

- The *key idea* of CART is based on **Recursive Partitioning:**
  - Repeatedly split the records into two parts so as to achieve maximum homogeneity within the new parts
- ***Gini index*** is a metric for classification tasks in CART.
  - Gini Index stores *sum of squared probabilities* of each class.

CART

- ***Gini index*** is a metric for classification tasks in CART.
  - It stores *sum of squared probabilities* of *each class*.
  - We can formulate it as,

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

where, c is the number of classes.

CART

- **Gini index** is illustrated as,

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

  - Gini(A) = 0 when all cases belong to same class
  - Max value when all classes are equally represented (= 0.50 in binary case)

CART

Recursive Partitioning

- Take all the data.
- Consider *all* possible values of *all* variables.
- Select the variable/value ($X=t_1$) that produces the **greatest "separation"** in the target.
  - ($X=t_1$) is called a "split".
- If $X< t_1$ then send the data to the **"left"**; otherwise, send data point to the **"right"**.
- Now repeat same process on these two "nodes"
  - Result into a "tree"
  - Note: CART only uses *binary* splits.

EXAMPLE (CART)

Example (CART)

	X				Y
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example (CART)

Outlook:

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

	X				Y
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Gini(Outlook=Sunny) $= 1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 0.48$

Gini(Outlook=Overcast) $= 1 - (4/4)^2 - (0/4)^2 = 0$

Gini(Outlook=Rain) $= 1 - (3/5)^2 - (2/5)^2 = 1 - 0.36 - 0.16 = 0.48$

Example (CART)

	X				Y
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Outlook:

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** for **outlook** feature.

Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48
= 0.171 + 0 + 0.171
= 0.342

Example (CART)

	X				Y
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Temperature:

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Gini(Temp=Hot) $= 1 - (2/4)^2 - (2/4)^2 = 0.5$

Gini(Temp=Cool)$= 1 - (3/4)^2 - (1/4)^2 = 1 - 0.5625 - 0.0625$

$= 0.375$

Gini(Temp=Mild) $= 1 - (4/6)^2 - (2/6)^2 = 1 - 0.444 - 0.111 = 0.445$

Example (CART)

Temperature:

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

	X				Y
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** for **temperature** feature.

Gini(Temp) = (4/14) x 0.5 + (4/14) x 0.375 + (6/14) x 0.445
= 0.142 + 0.107 + 0.190
= 0.439

Example (CART)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

X: Outlook, Temperature, Humidity, Wind; Y: PlayTennis

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Humidity:

Humidity	Yes	No	Number of instances
High	3	4	7
Normal	6	1	7

Gini(Humidity=High) $= 1 - (3/7)^2 - (4/7)^2 = 1 - 0.183 - 0.326$

$= 0.489$

Gini(Humidity=Normal) $= 1 - (6/7)^2 - (1/7)^2 = 1 - 0.734 - 0.02$

$= 0.244$

Example (CART)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Humidity:

Humidity	Yes	No	Number of instances
High	3	4	7
Normal	6	1	7

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** for **humidity** feature.

Gini(Humidity) = (7/14) x 0.489 + (7/14) x 0.244

= 0.367

Example (CART)

Wind:

Wind	Yes	No	Number of instances
Weak	6	2	8
Strong	3	3	6

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
	X				Y
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Gini(Wind=Weak) $= 1 - (6/8)^2 - (2/8)^2$

$= 1 - 0.5625 - 0.062 = 0.375$

Gini(Wind=Strong) $= 1 - (3/6)^2 - (3/6)^2$

$= 1 - 0.25 - 0.25 = 0.5$

Example (CART)

Wind:

Wind	Yes	No	Number of instances
Weak	6	2	8
Strong	3	3	6

	X				Y
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** for **wind** feature.

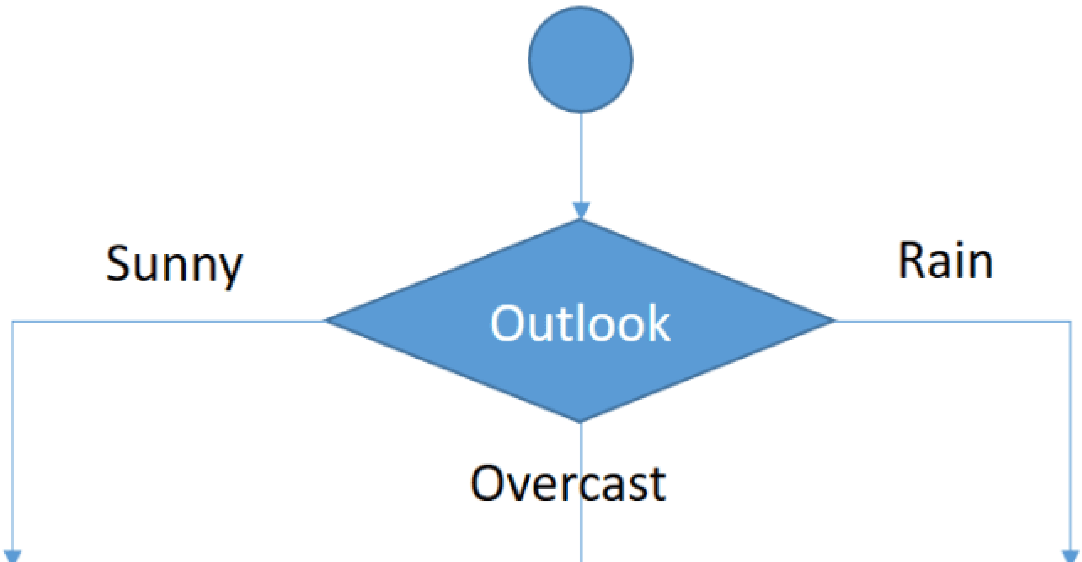
Gini(Wind) = (8/14) x 0.375 + (6/14) x 0.5

= 0.428

Example (CART)

Feature	Gini indices
Outlook	**0.342**
Temperature	0.439
Humidity	0.367
Wind	0.428

Example (CART)

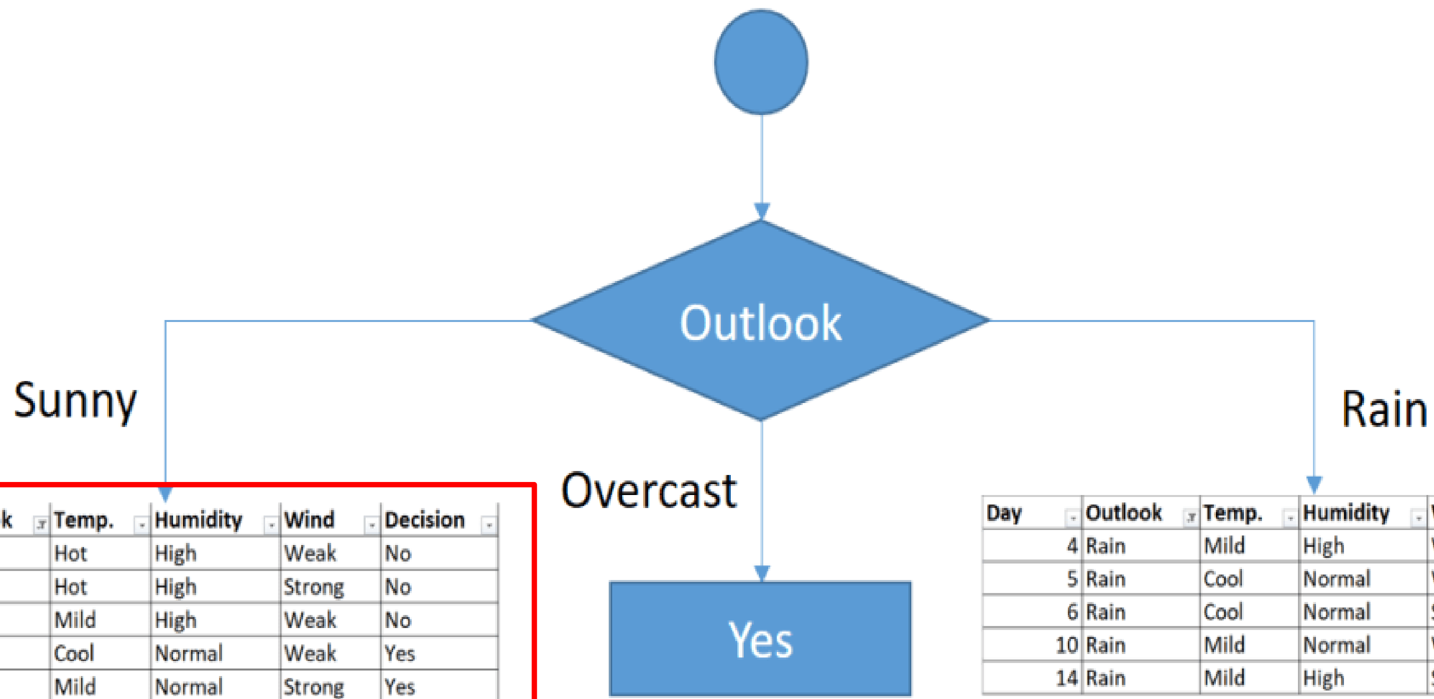


Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Example (CART)



Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Example (CART)

Sub-dataset for **sunny outlook**

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Example (CART)

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Temperature for sunny outlook:

Temperature	Yes	No	Number of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Gini(Outlook=Sunny and Temp.=Hot) $= 1 - (0/2)^2 - (2/2)^2 = 0$

Gini(Outlook=Sunny and Temp.=Cool) $= 1 - (1/1)^2 - (0/1)^2 = 0$

Gini(Outlook=Sunny and Temp.=Mild) $= 1 - (1/2)^2 - (1/2)^2$

$= 1 - 0.25 - 0.25 = 0.5$

Example (CART)

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Temperature for sunny outlook:

Temperature	Yes	No	Number of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** of **temperature** for **sunny outlook** feature.

Gini(Outlook=Sunny and Temp.) = (2/5)x0 + (1/5)x0 + (2/5)x0.5

= 0.2

Example (CART)

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Humidity for sunny outlook:

Humidity	Yes	No	Number of instances
High	0	3	3
Normal	2	0	2

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Gini(Outlook=Sunny and Humidity=High) $= 1 - (0/3)^2 - (3/3)^2$

$= 0$

Gini(Outlook=Sunny and Humidity=Normal) $= 1 - (2/2)^2 - (0/2)^2$

$= 0$

Example (CART)

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Humidity for sunny outlook:

Humidity	Yes	No	Number of instances
High	0	3	3
Normal	2	0	2

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** of **humidity** for **sunny outlook** feature.

Gini(Outlook=Sunny and Humidity) = (3/5)x0 + (2/5)x0

= 0

Example (CART)

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Wind for sunny outlook:

Wind	Yes	No	Number of instances
Weak	1	2	3
Strong	1	1	2

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Gini(Outlook=Sunny and Wind=Weak) $= 1 - (1/3)^2 - (2/3)^2$

$= 0.266$

Gini(Outlook=Sunny and Wind=Strong) $= 1 - (1/2)^2 - (1/2)^2$

$= 0.2$

CART

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Wind for sunny outlook:

Wind	Yes	No	Number of instances
Weak	1	2	3
Strong	1	1	2

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

We will calculate weighted sum of ***gini indices*** of **wind** for **sunny outlook** feature.

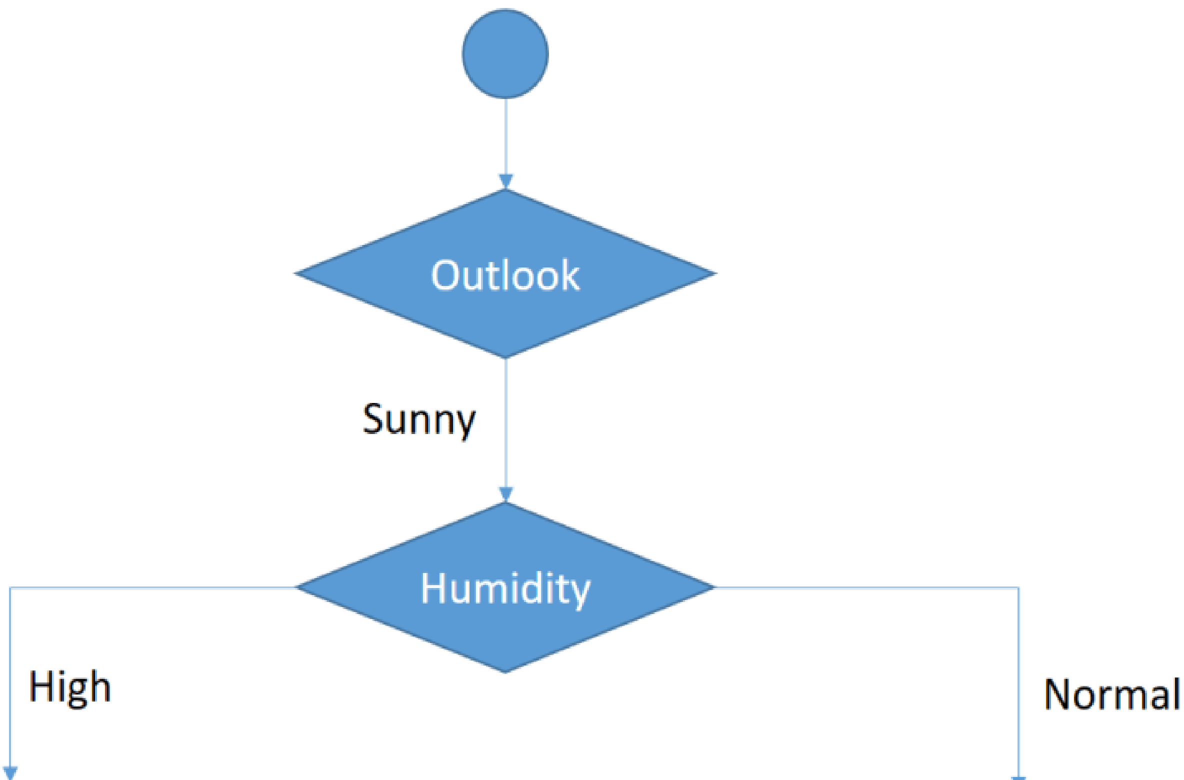
Gini(Outlook=Sunny and Wind) = (3/5)x0.266 + (2/5)x0.2

= 0.466

Example (CART)

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

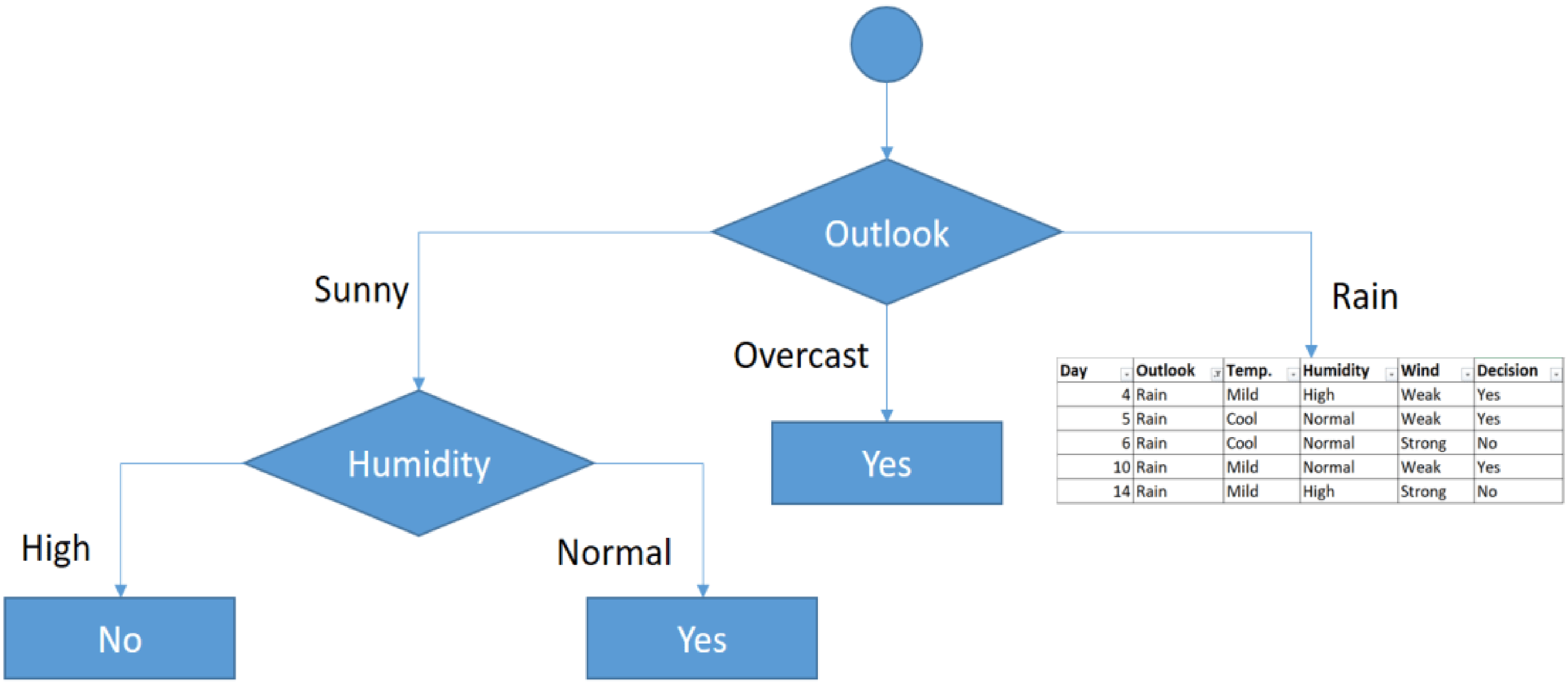
Example (CART)



Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

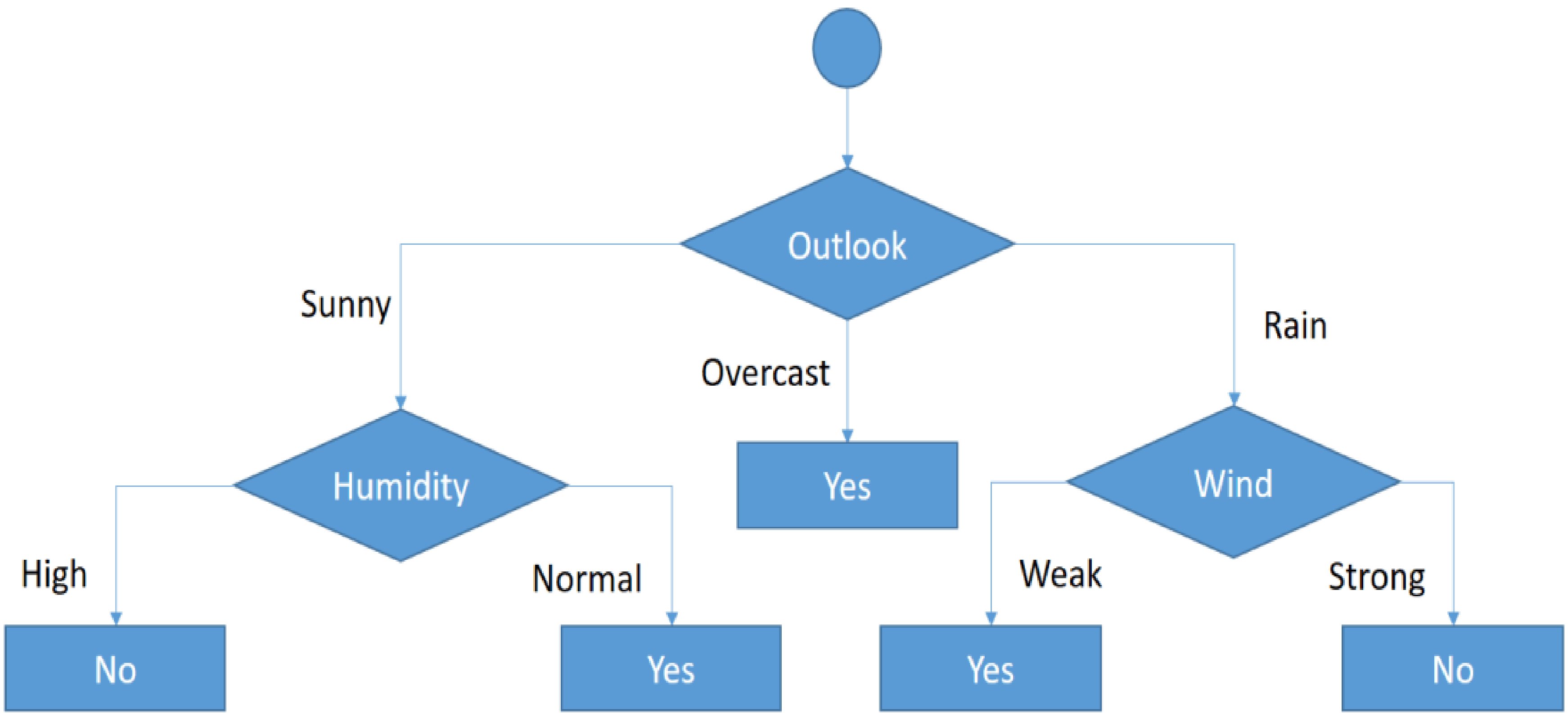
Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Example (CART)



Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Example (CART)



CART … Issues

- CART tree model can be overfit
  - CART can create a "branch" for every single training pixel.
  - Essentially modeling the noise in the training data, which is neither realistic nor practical
- Two options to control overfitting

  Stop Splitting Method: Stop growing the tree when further splitting the data *does not yield an improvement*

  Pruning: Build entire tree and then *remove branches that don't contribute much to accuracy* (considered better than "stop splitting").

Acknowledgement

Tom Mitchel, Russel & Norvig, Andrew Ng, Alpydin & Ch. Eick.