INSTITUT TEKNOLOGI BANDUNG

1920

Modul : Decision Tree Learning (DTL)

Variable (Attribute) Types

Source: DataMining Concepts and Techniques
by Jiawei Han, Micheline Kamber, Jian Pei

Nur ULFA Maulidevi

KK IF - Teknik Informatika- STEI ITB

Pembelajaran Mesin
(Machine Learning)

Numeric

Quantitative (measurable quantity) → Integer or Real Values

Interval Scaled (equal-size units, have order)

- Temperature
- Calendar dates

Ratio-Scaled (inherent zero point, a value can be multiple of another value)

- Years of experience
- Weight
- Number of words

Nominal/ Categorical

Symbols or Name of things

Do not have meaningful order

Can represent names with numbers → not to be used quantitatively

- Occupation Type
- Zip Code
- Student_ID

Binary

Only two categories or states (Boolean if states: true and false)

Symmetric: equally valuable and carry the same weight

Asymmetric: not equally important

- Gender
- HIV Positive/ negative

```
red,    green,  blue
1,      0,      0
0,      1,      0
0,      0,      1
```

Ordinal

Have meaningful order/ ranking of possible values

magnitude between successive values is not known

useful for registering subjective assessments of qualities

May be obtained from discretization of numeric quantities

- Drink Size: small, medium, large
- Customer Satisfaction: 1, 2, 3, 4, 5
- Grades: E,D,C,BC,B,AB,A

Discrete vs Continuous

has a finite set of values:
Drink size, Age, Medical test,

has a countably infinite set of values:
Customer ID, Zip code

Discrete

≠

Continuous

Modul : Decision Tree Learning (DTL)

What, Why, and When

Source: Machine Learning, Tom Mitchell

Nur ULFA Maulidevi

KK IF - Teknik Informatika- STEI ITB

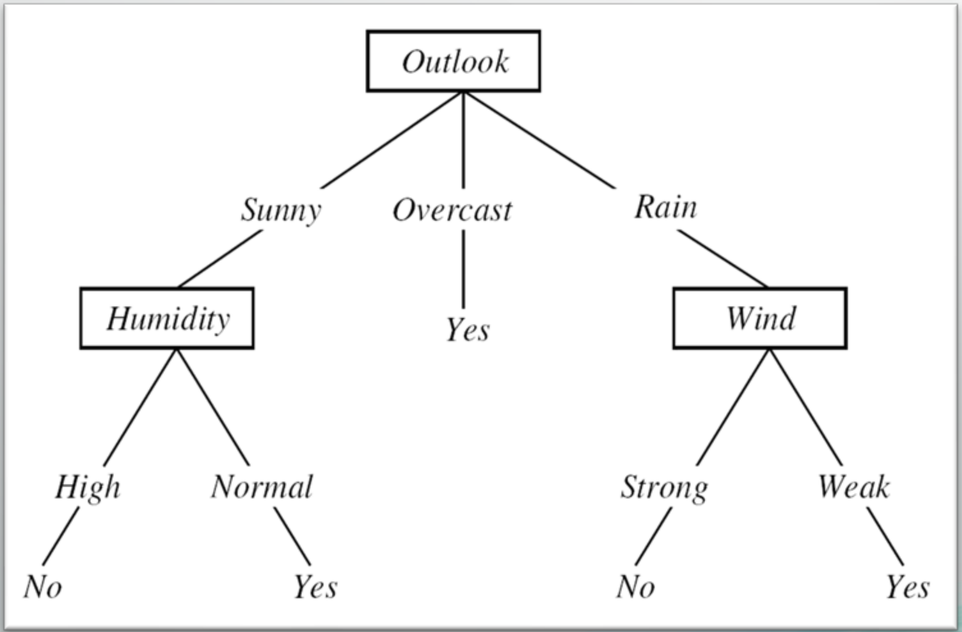
Pembelajaran Mesin
(Machine Learning)

WHAT is DTL

Method for approximating discrete-valued target functions

Represented by Decision Tree

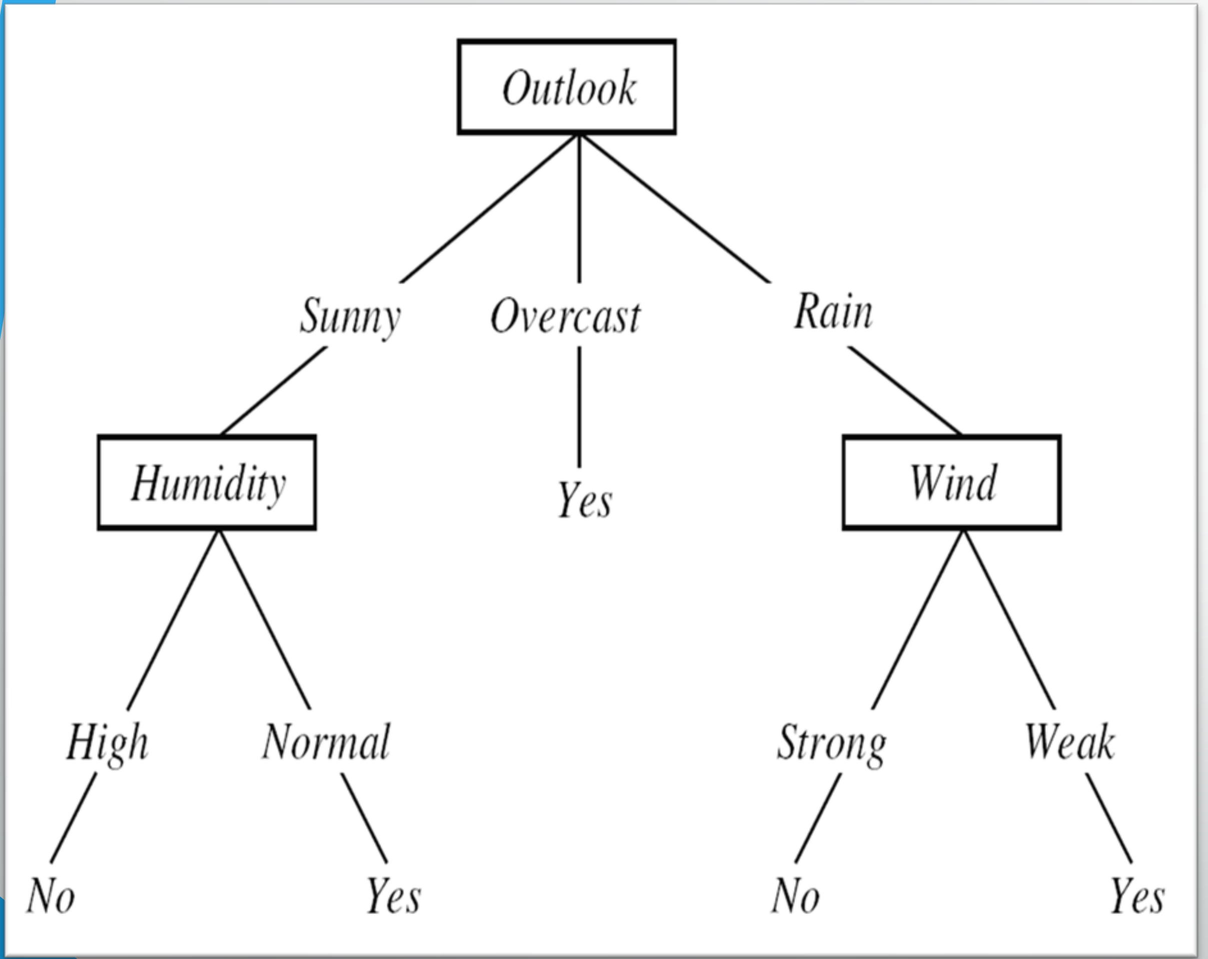
Can be represented as set of if-then rules



EDUNEX ITB

DTL Representation

constraint tu nilai dari atribut yang dipilih



- Each internal node represents test of an attributes
- Branch descending from the node corresponds to one possible value
- Leaf/ terminal nodes represents classification result

Represent a disjunction of conjunctions of constraints on the attribute values of instances

(outlook=sunny ∧ humidity=normal)
∨ (outlook=overcast)
∨ (outlook=rain ∧ wind=weak)

WHY DTL

Method for approximating discrete-valued target functions

Represented by Decision Tree or If-Then Rules

pola data bisa dibaca lgsg sm manusia

Robust to Noisy Data

noisy data tdk berpengaruh krn data bsr

Capable of learning disjunctive expression

- **Popular of inductive inference algorithms**
- **Have been successfully applied to a broad range of tasks: learning to diagnose medical cases, learning to assess credit risk of loan applicants**

When we use DTL → Appropriate Problems for DTL

data berlabel

Instances are represented by attribute-value pairs → **For continuous attribute: discretization**

The target function has discrete output values

Disjunctive descriptions may be required

The training data may contain errors

The training data may contain missing attribute values

Modul : Decision Tree Learning (DTL)

Basic DTL Algorithm (ID3)

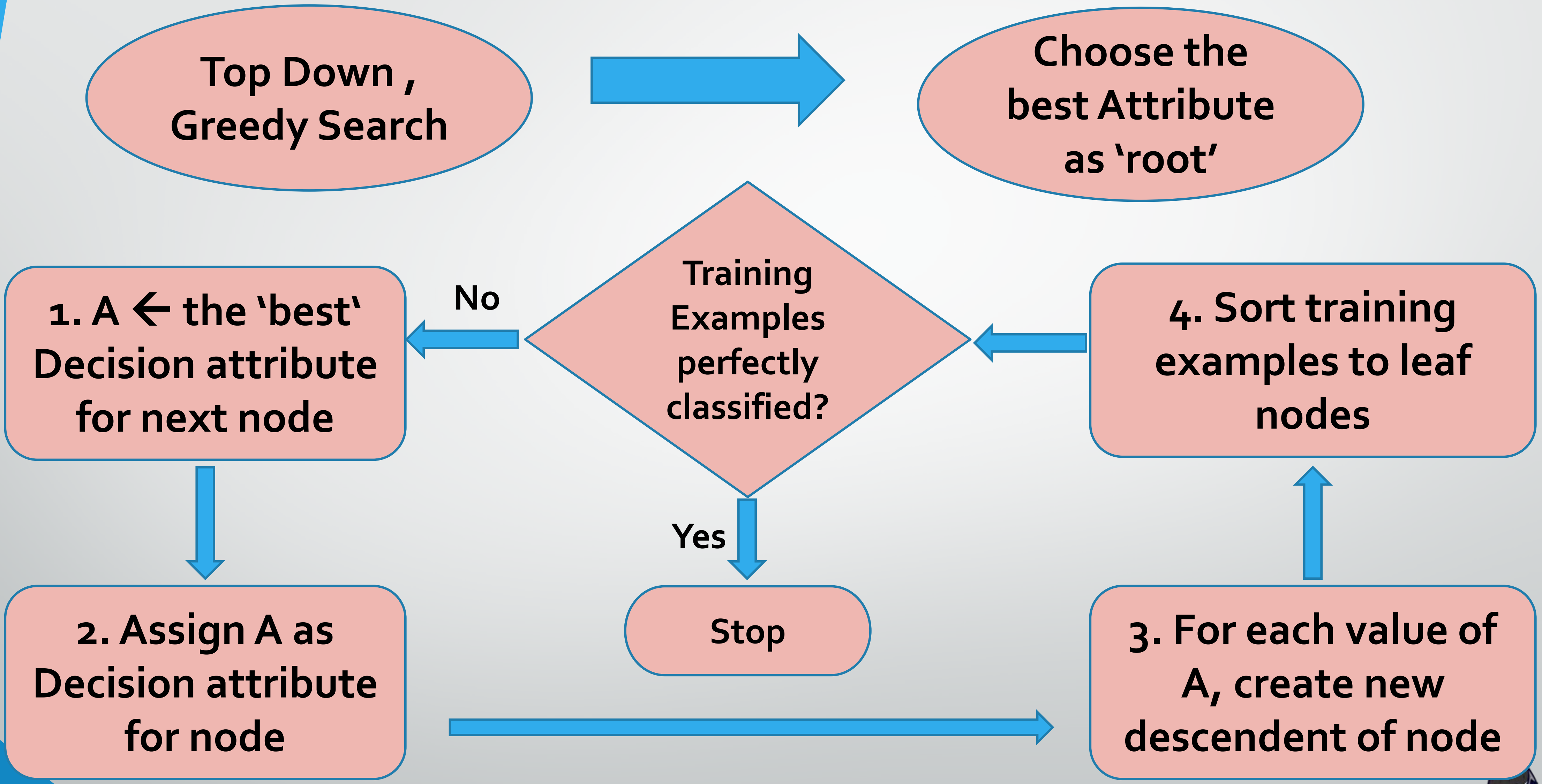
Source: Machine Learning, Tom Mitchell

Nur ULFA Maulidevi

KK IF - Teknik Informatika- STEI ITB

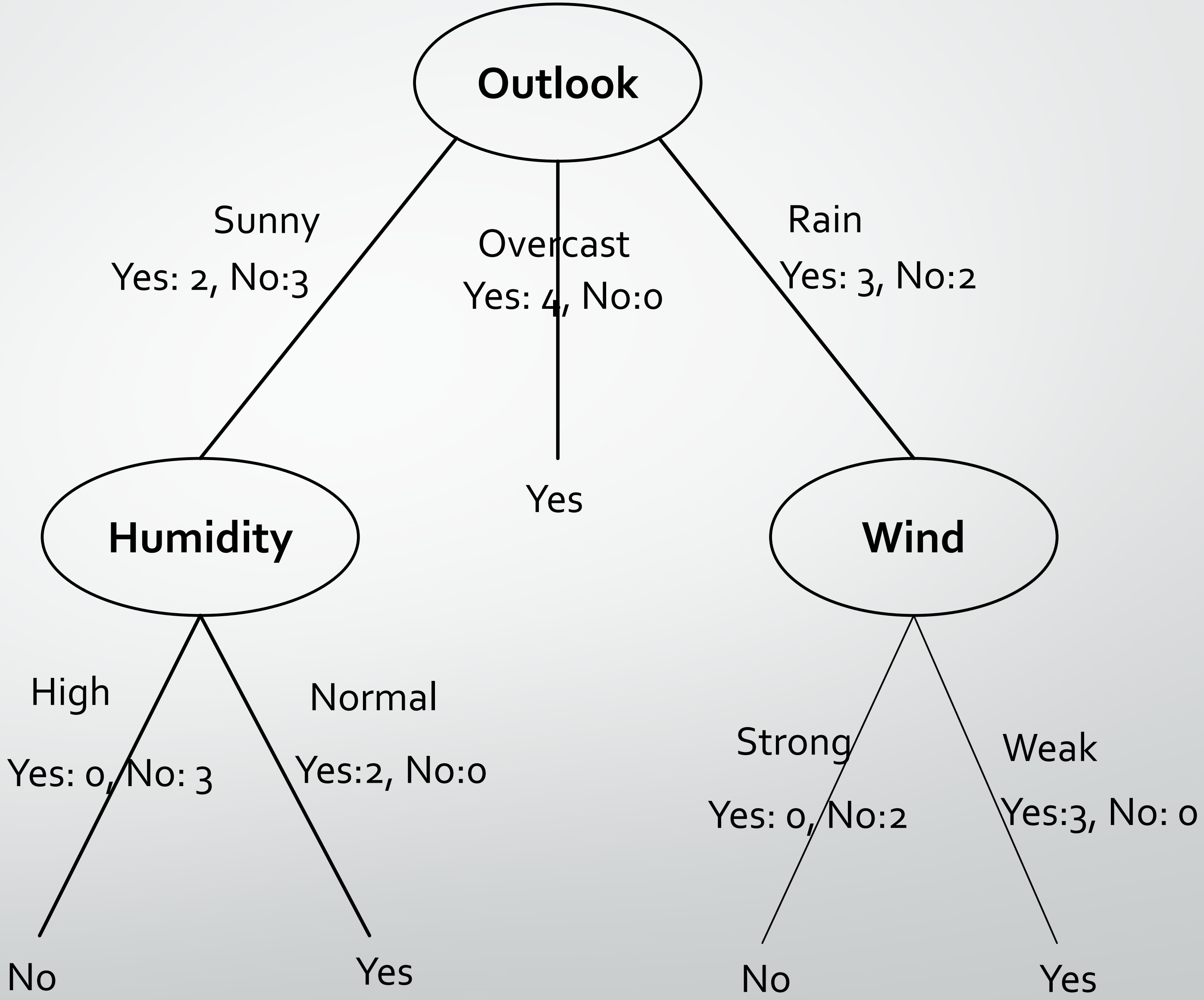
Pembelajaran Mesin
(Machine Learning)

ID3 Algorithm



Example

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

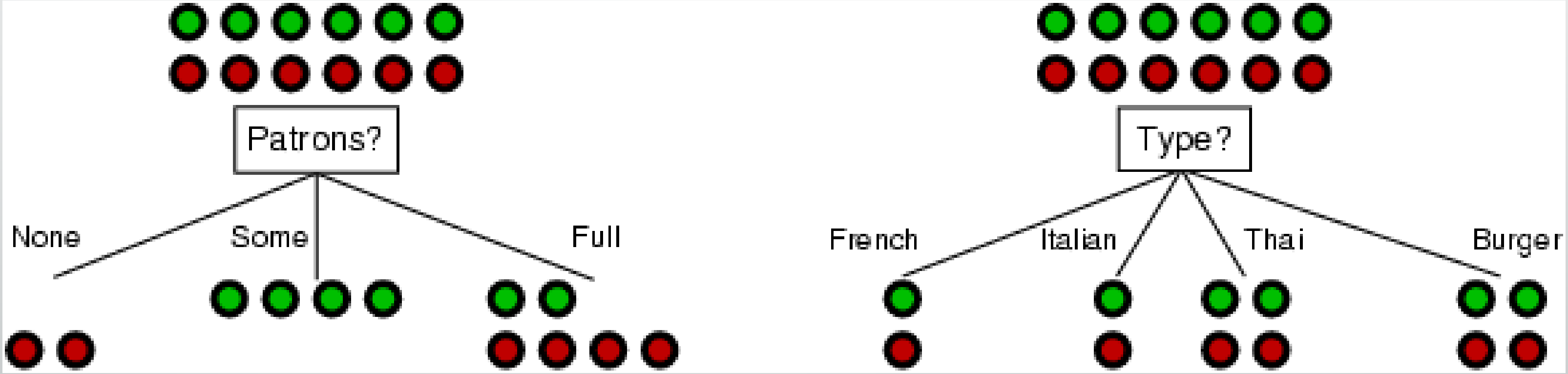


Best Attribute

"Ideally" can classify the training examples into subsets, which has the same class

→

Information Gain



Modul : Decision Tree Learning (DTL)

Information Gain

Source: Machine Learning, Tom Mitchell

Nur ULFA Maulidevi

KK IF - Teknik Informatika- STEI ITB

Pembelajaran Mesin
(Machine Learning)

Entropy (Information Content)

Measurement in Information Theory → impurity of an arbitrary collection of samples

S = set of training examples
Entropy: the minimum number of bits of information needed to encode the classification of an arbitrary member of S

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

- S: set of training examples
- c: number of classes
- p_i: proportion of S belonging to class i

Entropy for S with 2 values/ classes

S contains positive examples and negative examples

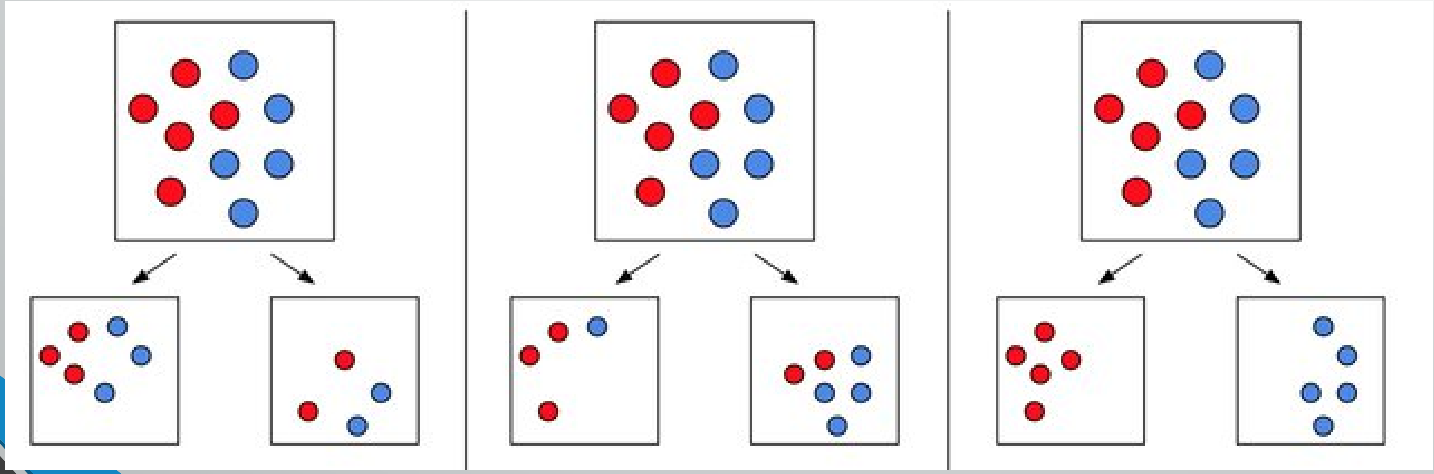
$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Entropy 0: all examples belong to a class (no surprises, no message need be sent)

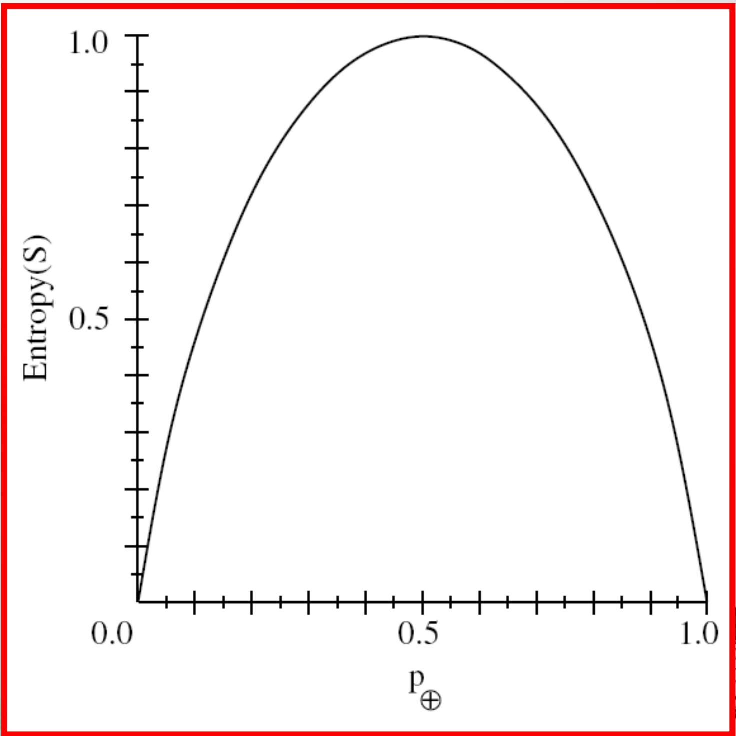
Entropy 1: $p_+ = p_-$

(1 bit is required to indicate the class of the drawn example)

I(5/10,5/10) = 1



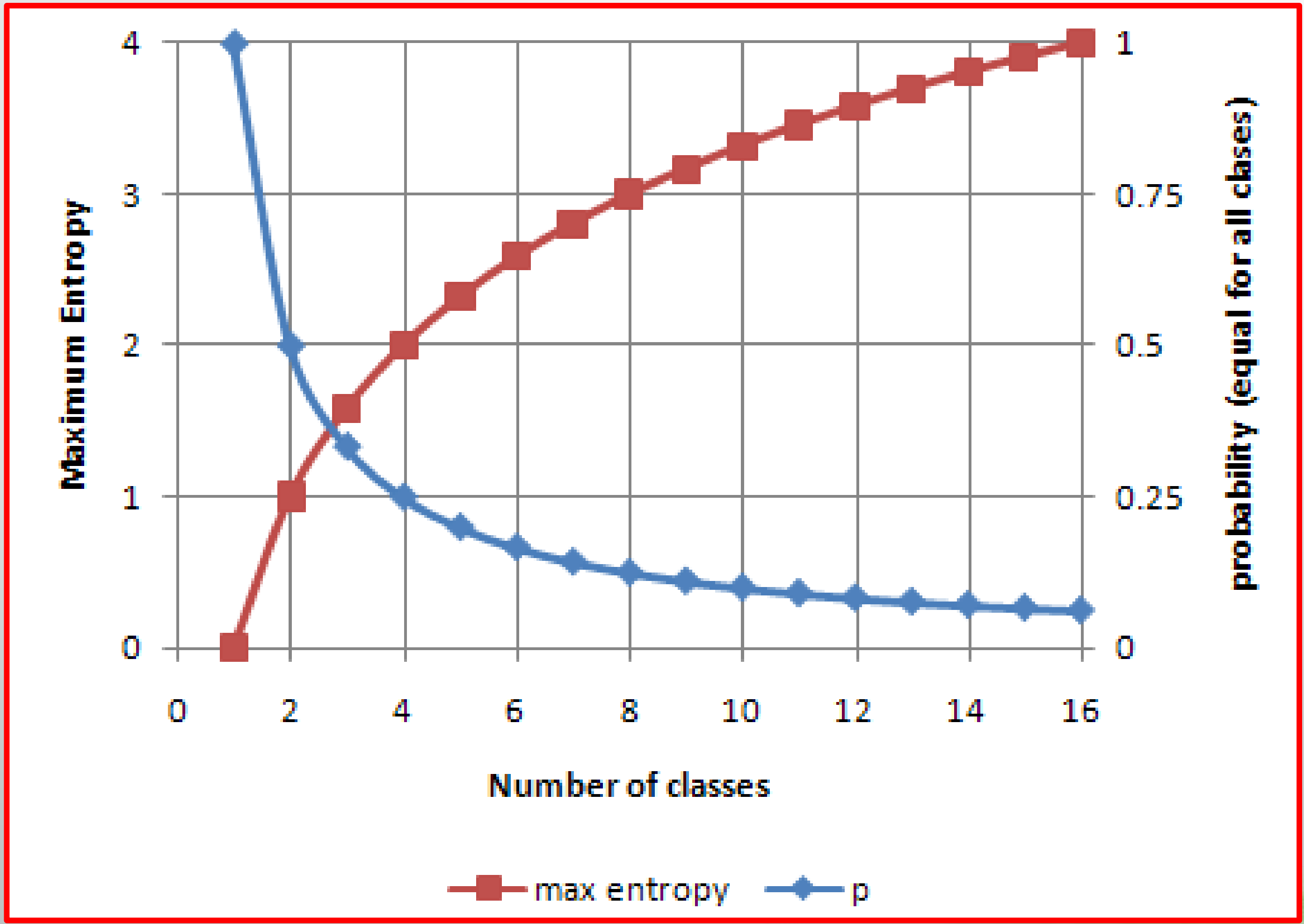
I(3/6,3/6) = 1 I(2/4,2/4) = 1 I(3/4,1/4)=0.81 I(2/6,4/6)=0.92 I(5/5,0/5)=0 I(0/5,5/5)=0



https://www.researchgate.net/publication/320384121_Textklassification_-_Decision_Trees_Random_Forests/figures?lo=1

Entropy for S with n values/ classes

Target attribute has n values: entropy can be as large as $\log_2 n$



Information Gain

Gain(S,A) = expected reduction of entropy due to sorting A
S: set of training examples; A: an attribute

Find A which has maximum Gain(S,A)

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

EDUNEX ITB

Example

Find the 'best' Attribute:
max(Gain(S,Outlook),
Gain(S,Temperature),
Gain(S,Humidity),
Gain(S,Wind))

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

$$Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.940$$

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in value\ of\ Outlook} \frac{S_v}{S} Entropy(S_v)$$

$$= 0.940 - [\frac{5}{14} Entropy(Sunny) + \frac{4}{14} Entropy(Overcast) + \frac{5}{14} Entropy(Rain)]$$

$$= 0.940 - [(\frac{5}{14}(-\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5})) + (\frac{4}{14}(-\frac{4}{4} log_2 \frac{4}{4} - \frac{0}{4} log_2 \frac{0}{4})) + (\frac{5}{14}(-\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5}))]$$

$$= 0.246$$

EDUNEX ITB

Example

Find the 'best' Attribute:
max(Gain(S,Outlook),
Gain(S,Temperature),
Gain(S,Humidity),
Gain(S,Wind))

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

$$Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.940$$

$$Gain(S, Humidity) = Entropy(S) - \sum_{v \in value\ of\ Humidity} \frac{S_v}{S} Entropy(S_v)$$

$$= 0.940 - [\frac{7}{14}Entropy(High) + \frac{7}{14}Entropy(Normal)]$$

$$= 0.940 - [(\frac{7}{14}(-\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7})) + (\frac{7}{14}(-\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7}))]$$

$$= 0.151$$

EDUNEX ITB

Example

Find the 'best' Attribute:
max(Gain(S,Outlook),
Gain(S,Temperature),
Gain(S,Humidity),
Gain(S,Wind))

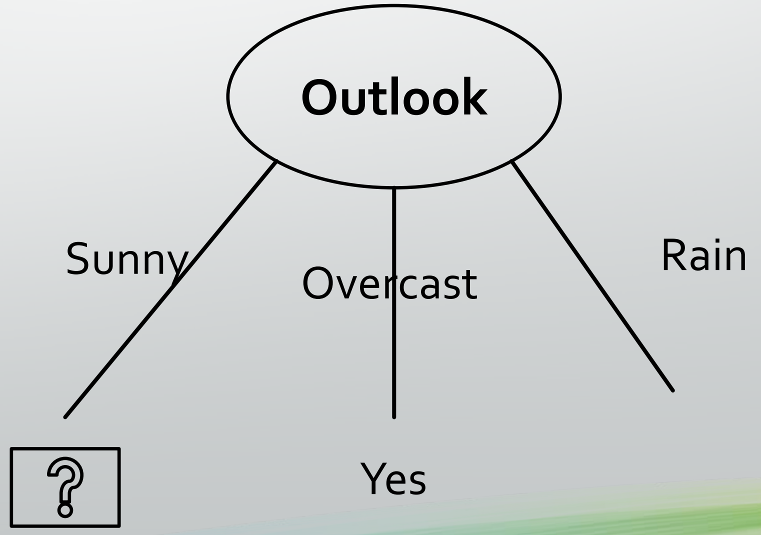
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal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

$$Gain(S, Outlook) = 0.246$$

$$Gain(S, Humidity) = 0.151$$

$$Gain(S, Temperature) = 0.029$$

$$Gain(S, Wind) = 0.048$$



EDUNEX ITB

Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

Find the 'best' Attribute:
max(Gain(S,Temperature),
Gain(S,Humidity),
Gain(S,Wind))

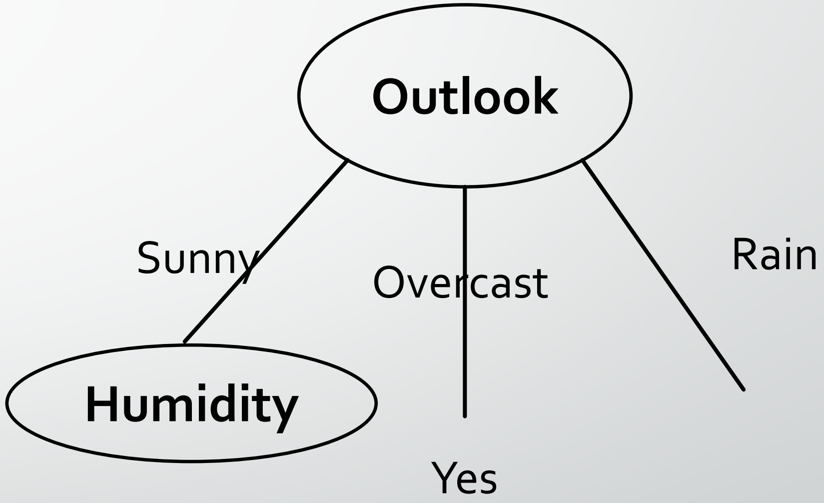
$$Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$Gain(S, Humidity) = Entropy(S) - \sum_{v \in value\ of\ Humidity} \frac{S_v}{S} Entropy(S_v)$$

$$= 0.971 - \frac{3}{5}Entropy(High) - \frac{2}{5}Entropy(Normal)$$

$$= 0.971 - [(\frac{3}{5}(-\frac{0}{3}log_2\frac{0}{3} - \frac{3}{3}log_2\frac{3}{3}) + (\frac{2}{5}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}))]$$

$$= 0.971$$



$$Gain(S, Humidity) = 0.971$$

$$Gain(S, Temperature) = 0.571$$

$$Gain(S, Wind) = 0.020$$

EDUNEX ITB

Exercise (Class: Aktivitas)

No	Deadline?	Ada Hangout?	Malas?	Aktivitas
1	Urgent	ya	Ya	Kumpul-kumpul
2	Urgent	Tidak	Ya	Belajar
3	Dekat	ya	Ya	Kumpul-kumpul
4	Tidak ada	Ya	Tidak	Kumpul-kumpul
5	Tidak ada	Tidak	Ya	Jalan-jalan ke mall
6	Tidak ada	Ya	Tidak	Kumpul-kumpul
7	Dekat	Tidak	Tidak	Belajar
8	Dekat	Tidak	Ya	Nonton TV
9	Dekat	Ya	Ya	Kumpul-kumpul
10	Urgent	Tidak	Tidak	Belajar

THANK YOU

INSTITUT TEKNOLOGI BANDUNG

1920