

# Modul : Issues in Decision Tree Learning (DTL)

## Missing Attribute Value

Nur ULFA Maulidevi

KK IF - Teknik Informatika- STEI ITB

Pembelajaran Mesin  
(Machine Learning)



# Issues in DTL

Overfitting  
training  
data

Continuous  
-valued  
attribute

Handling  
attributes  
with differing  
costs

Handling  
missing  
attribute  
value

Alternative  
measures for  
selecting  
attributes



# Alternative Strategies

Assign it with the most common value at node  $n$  among other examples

Assign it with the most common value at node  $n$  that have classification  $c(x)$

Assign probability  $p_i$  to each possible value  $v_i$  of  $A$  (used in C4.5)

The probability  
Can be used for  
classifying  
a new  
instance  
with  
missing  
value

Gain( $S, A$ ) only consider the fraction of training examples with known value  
 $\text{Gain}(S, A) = 10/11 * (\text{Entropy}(S) - [\sum \text{proportion} * \text{entropy\_of\_known\_value}])$

$v_1 = 1$ , 6 known examples;  $v_2 = 0$ , 4 known examples, 1 example with missing value of attr  $A$   
 $p_i = 6/10$  added to  $v_1$  ;  $p_2 = 4/10$  added to  $v_2 \rightarrow$  for splitting

# Missing Value as Separate Value

Denoted "?" → Null Value  
In C4.X

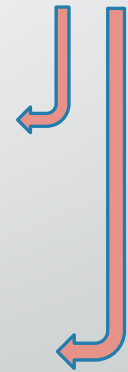
Not  
Appropriate  
when:



Values are missing due to different reasons

blood sugar value could be missing when it is very high or very low

field IsPregnant missing for a male patient should be treated differently (no) than for a female patient of age 25 (unknown)





---

# THANK YOU



