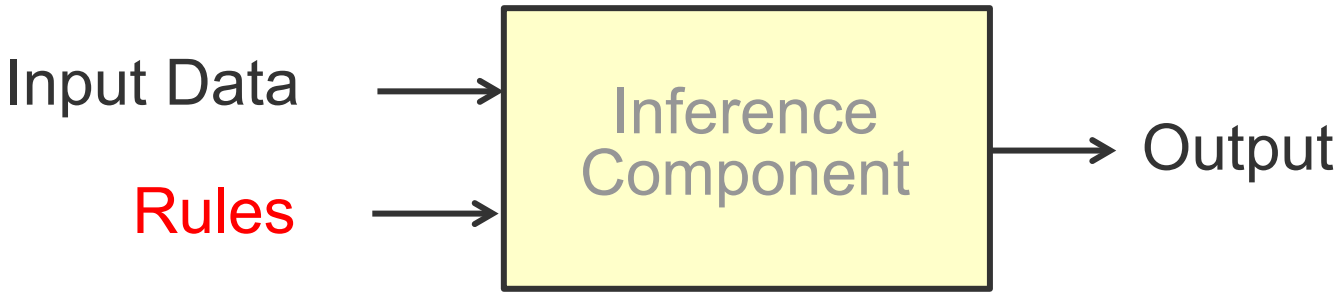
University of Applied Sciences and Arts Northwestern Switzerland
School of Business

Symbolic Machine Learning: Learning Decision Trees

Knut Hinkelmann

Symbolic Machine Learning vs. Knowledge-based Systems

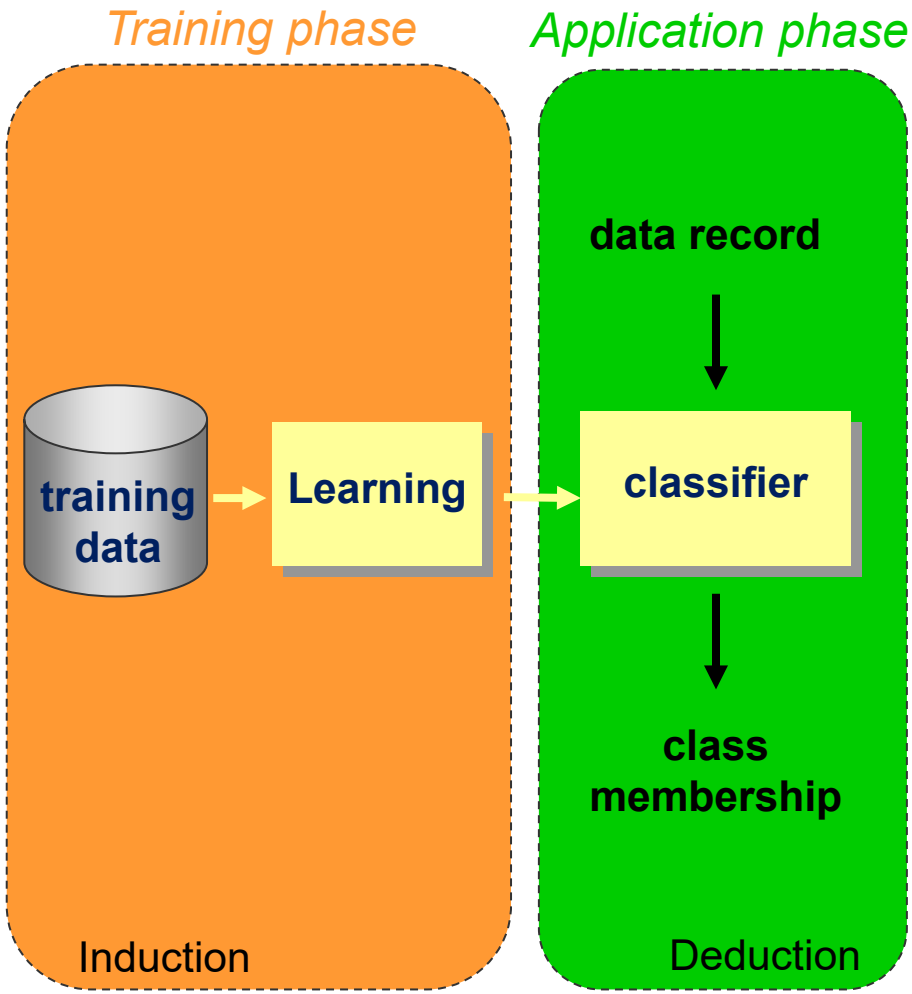
Knowledge-based System



Machine Learning

Source: Vibhav Gogate, UT Dallas

Training and Application Phase



- **Training:** Learning the classification criteria
  - Given: sample set of training data records
  - Result: Decision logic to determine class from values of input attributes
- **Application:** Classification
  - Assign a class to previously unseen records of input data

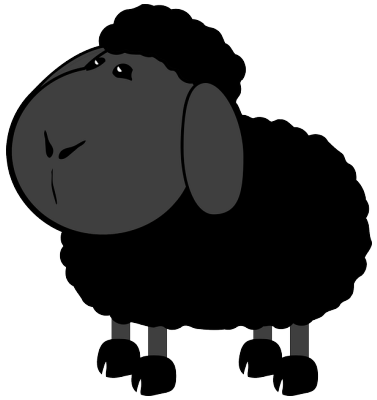
Example

Given a number of data sets, which provide observation which weather has been good for playing tennis in the past.

Element	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Tennis*
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Challenge: Can we use this data to determine in advance, whether we should go for playing tennis
- Naive approach: Each observed data set represents a rule
  - Problem: Not all cases are represented
  - Example: What happens if the outlook is «Rain», the Temperature is «Hot» and the wind is «Weak»?
- **Machine Learning:**
  Generalize the data to a set of rules which are applicable also in cases that are not covered by the data

The Problem of Generalization



A sociologist, an economist, a physicist and a mathematician go by train to Scotland. They look out of the window and see a black sheep.

Sociologist: „In Scotland the sheeps are black“

Economist: „Wrong, in Scotland there are black sheeps“

Physicist: „Wrong, in Scotland there is at least one black sheep.“

Mathematician: „Still wrong. In Scotland there is a least on sheep that is black on a least one side“

Predictive Model for Classification

Training data

input			*class*
…	…	…	…
…	…	…	…
…	…	…	…

Learning

Decision logic

Playing Tennis				
	Outlook	Humidity	Wind	Tennis
	Sunny,Overcast, Rain	*High, Normal*	*Strong,Weak*	*Yes, No*
1	Sunny	High		No
2	Sunny	Normal		Yes
3	Overcast			Yes
4	Rain		Strong	No
5	Rain		Weak	Yes

- Given a collection of training data (*training set*)
  - Each data record consists of *attributes*, one of the attributes is the *class*
    - The class is the dependent attribute
    - The other attributes are the independent attributes
- Learning: Find a *model* for the class attribute as a function of the values of the other attributes.
  - Goal: to assign a class to **previously unseen records** as accurately as possible.
- Generalisation of data if training set does not cover all possible cases or data are too specific
  - → Induction

Example

The dependent variable „Tennis“ determines if the weather is good for tennis („Yes“) or not („No“).

Training Data

Element	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Tennis*
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Induction generalizes the data set → prediction of future case

Playing Tennis				
	Outlook	Humidity	Wind	Tennis
	Sunny,Overcast, Rain	*High, Normal*	*Strong,Weak*	*Yes, No*
1	Sunny	High		No
2	Sunny	Normal		Yes
3	Overcast			Yes
4	Rain		Strong	No
5	Rain		Weak	Yes

The result of the induction algorithms classifies the data with only three of the four attributes into the classes „Yes“ and „No“.

Discussion

What is the difference between the table with the Training Data and the Decision Table?

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Playing Tennis				
	Outlook	Humidity	Wind	Tennis
	Sunny,Overcast, Rain	*High, Normal*	*Strong,Weak*	*Yes, No*
1	Sunny	High		No
2	Sunny	Normal		Yes
3	Overcast			Yes
4	Rain		Strong	No
5	Rain		Weak	Yes

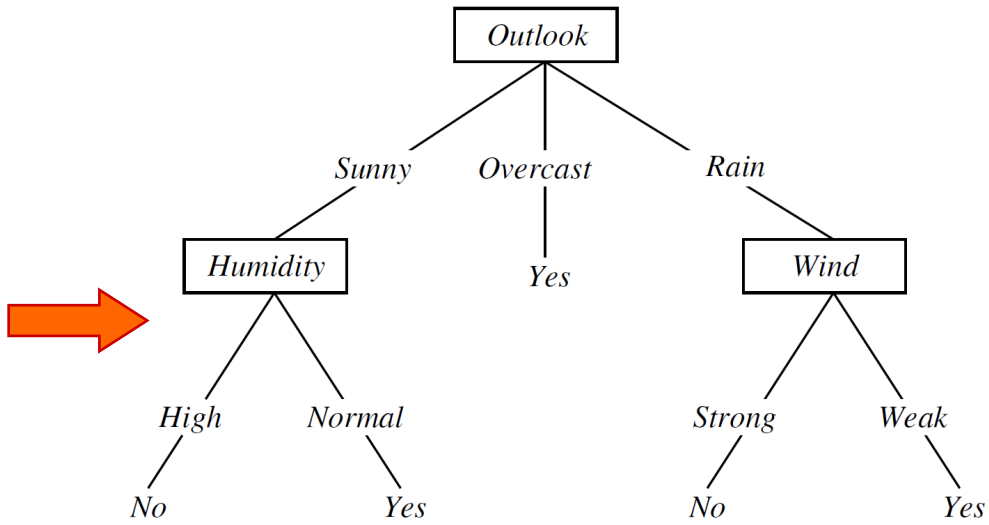
Training Data vs. Decision Tables (Rules)

- Training Data are …
  - … incomplete: only a subset of all possible situations
  - … too specific: they contain input variables, which are not necessary to determin the output
- Rule set shall be general, i.e. allow decisions/ predictions for unknown situations
  - Rules only consider combinations of input values, which are necessary to determine the output
  - As a consequence, the decision table does not contain variables, which are not necessary at all
    (e.g. playing tennis does not depend on the temperature)

Learning Decision Trees

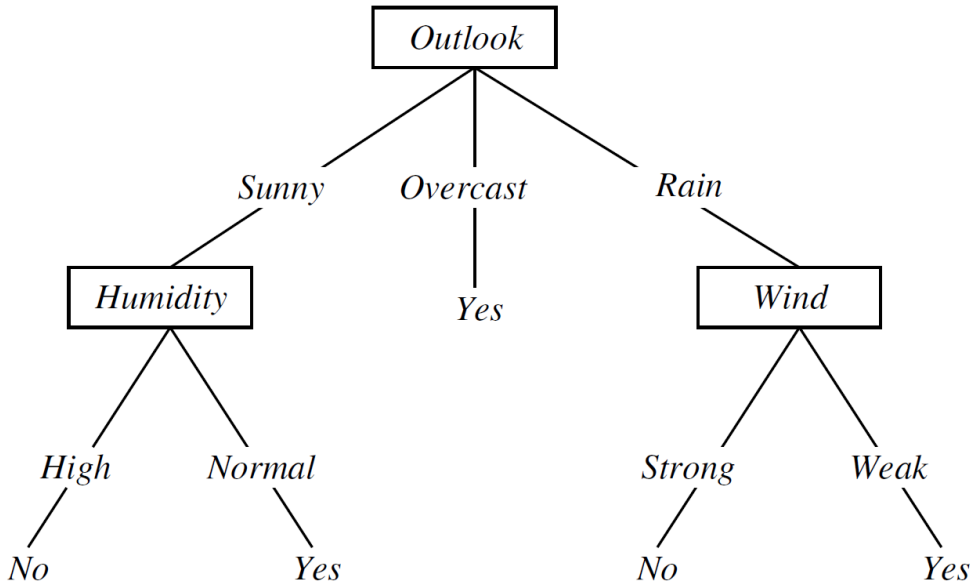
Training Data

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



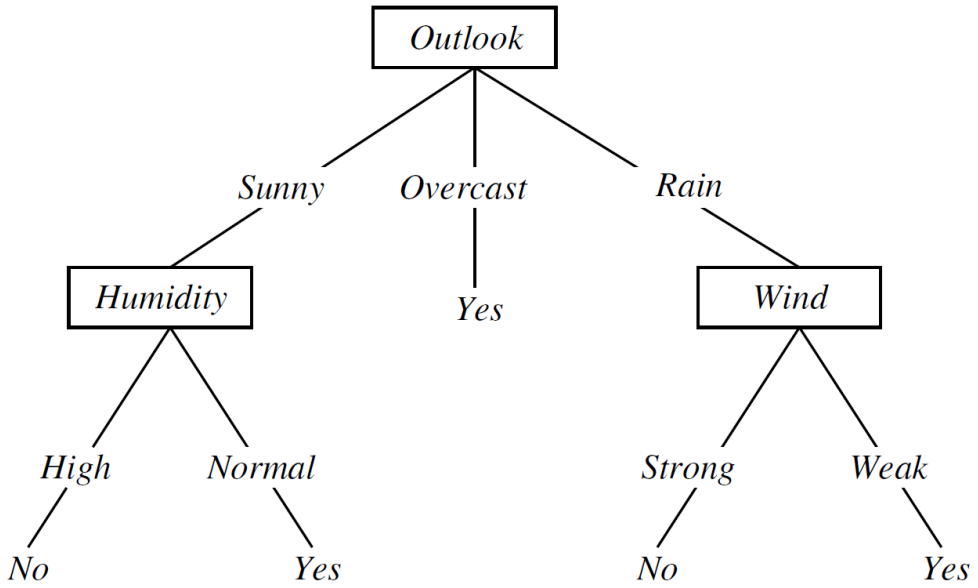
Decision Trees

Example: Decision tree for playing tennis



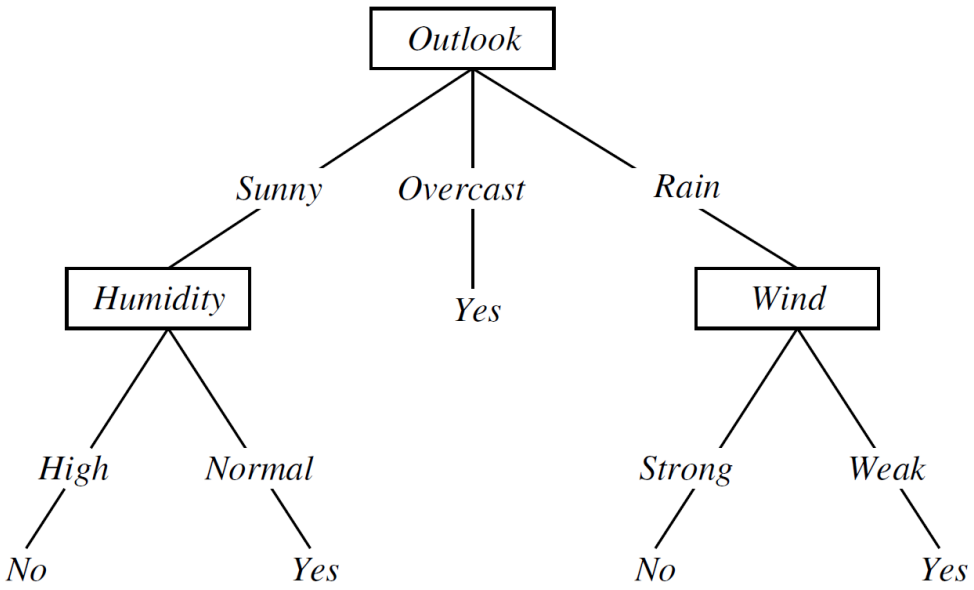
- Decision trees are primarily used for classification
- Decision trees represent classification rules
- Decision tree representation:
  - Each internal node tests an attribute
  - Each branch corresponds to attribute value
  - Each leaf node assigns a classification
- Decision trees classify instances by sorting them down the tree from the root to some leaf node,

Decision Trees represent Rules



- Each path from root to a leaf is a rule
- Each path/rule is a conjunction of attribute tests:
  - **IF** Outlook = Sunny AND Humidity = High **THEN** No
  - **IF** Outlook = Sunny AND Humidity = Normal **THEN** Yes
  - **IF** Outlook = Overcast **THEN** Yes
  - **IF** Outlook = Rain AND Wind = Strong **THEN** No
  - **IF** Outlook = Rain AND Wind = Weak **THEN** Yes

Decision Trees represent Rules

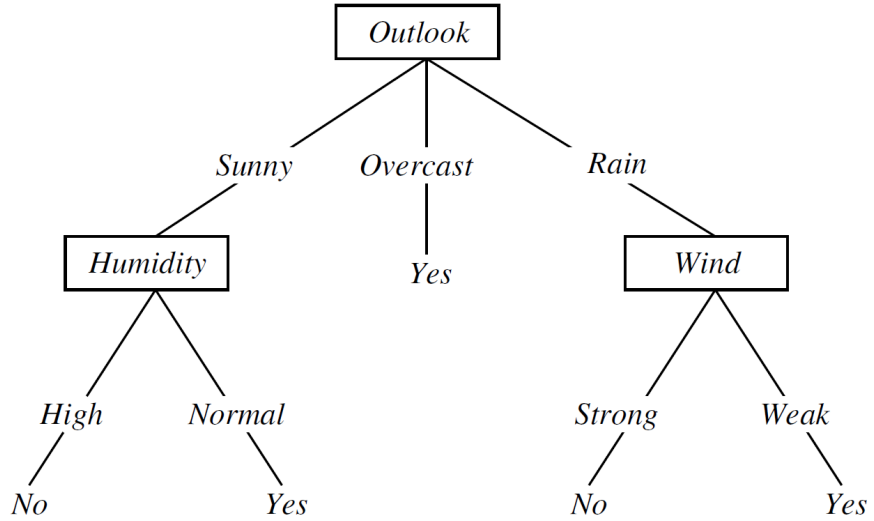


- If the classes are boolean, a path can be regarded as a conjunction of attribute tests.
- The tree itself is a disjunction of these conjunctions

$$(\text{Outlook} = \text{Sunny} \wedge \text{Humidity} = \text{Normal})$$
$$\vee$$
$$(\text{Outlook} = \text{Overcast})$$
$$\vee$$
$$(\text{Outlook} = \text{Rain} \wedge \text{Wind} = \text{Weak})$$

Decision Tree – Decision Table

The decision tree can be represented as a decision table.



Playing Tennis

	Outlook	Humidity	Wind	Tennis
	Sunny,Overcast, Rain	*High, Normal*	*Strong,Weak*	*Yes, No*
1	Sunny	High		No
2	Sunny	Normal		Yes
3	Overcast			Yes
4	Rain		Strong	No
5	Rain		Weak	Yes

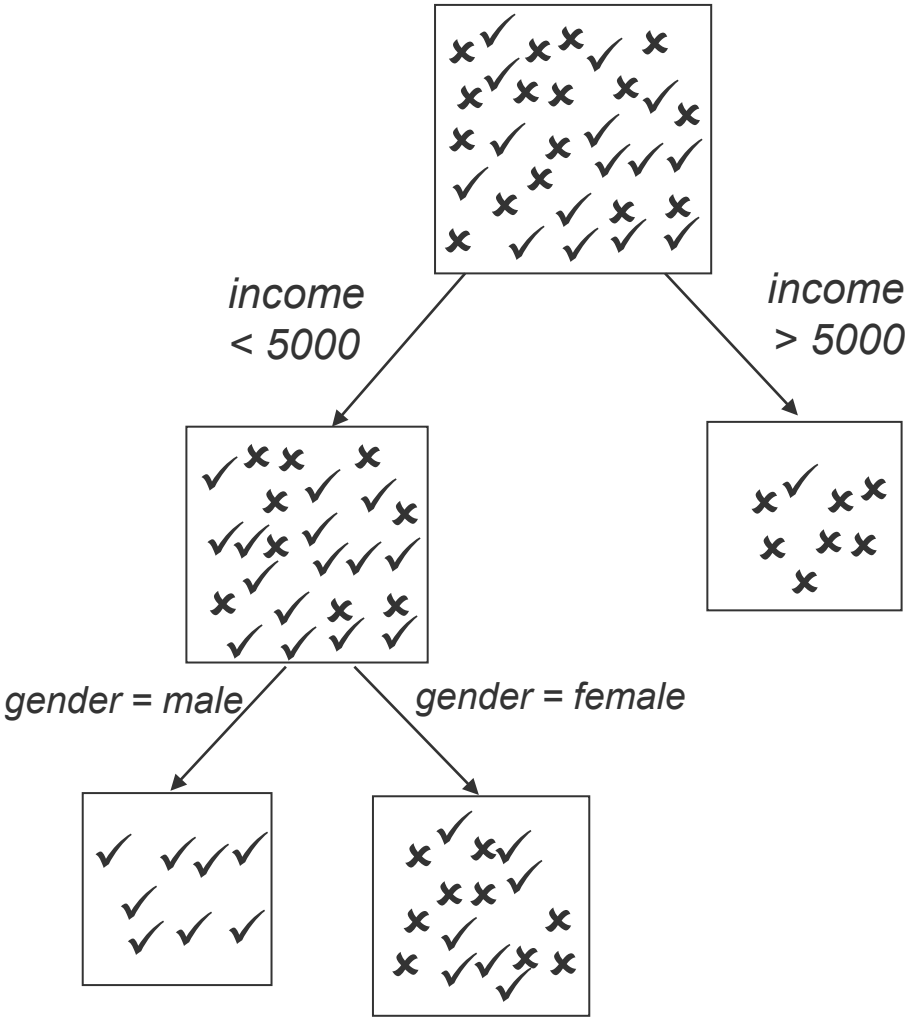
Learning a Decision Tree

Principle: From heterogenous classes to homogenous classes

Heuristic approach:

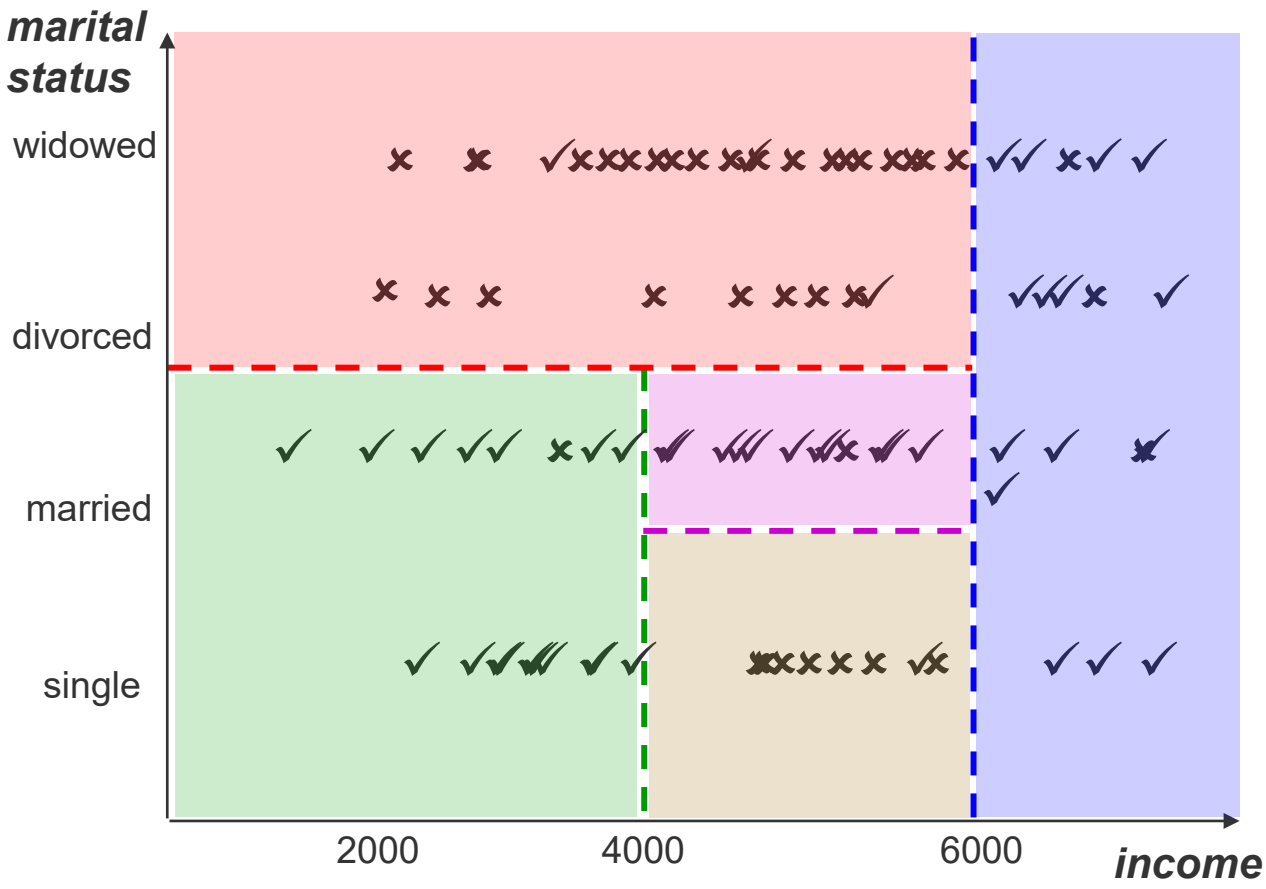
- Start with the full set of elements
- Extend the tree step by step with new decision criteria: Each decision should lead to more homogenous subsets
- Stop, if the desired homogenity is achieved

This approach is efficient, but does not necessariy find the best classifying tree.



Creation of Decision Trees

Each decision divides the area in sections



IF income > 6000
THEN accept

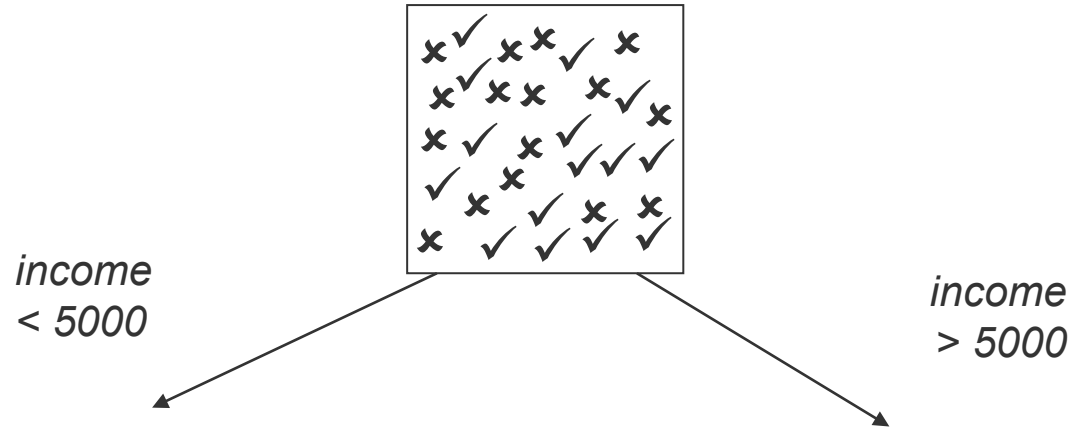
IF income <= 6000 and marital status = widowed or marital status = divorced
THEN reject

IF income <= 4000 and marital status = single or marital status = married
THEN accept

IF income > 4000 and income <= 6000 and marital status = married
THEN accept

IF income > 4000 and income <= 6000 and marital status = single
THEN reject

Determine how to split the Records in a Decision Tree



- **Attribute selection**
  - Which **attributes** separate best in which order?
    - e.g. income before marital status
- **Test condition**
  - Which **values** separate best?
    - Discrete: select value, e.g. single or married
    - Number: determine splitting number, e.g. income < 5000

Types of Data

- Discrete: final number of possible values
  - Examples: marital status, gender
  - Splitting: selection of values or groups of values
- Numeric infinite number of values on which an order is defined
  - Examples: age, income
  - Splitting: determine interval boundaries

For which kind of attributes is splitting easier?

Heuristic Induction: Principle

Learning a Decision Tree

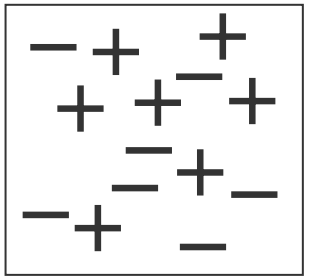
- Calculate for each attribute, how ***good*** it classifies the elements of the training set
- Classify with the ***best*** attribute
- *Repeat* for each subtree the first two steps
- Stop this recursive process as soon as a ***termination condition*** is satisfied

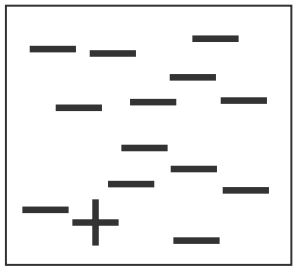
Generating Decision Trees

- ID3 is a basic decision learning algorithm.
- It recursively selects test attributes and begins with the question "*Which attribute should be tested at the root of the tree?* “
- ID3 selects the attribute with the highest
  - **Information Gain**
    (this is the attribute with reduces entropy the most)
- To calculate the information gain of an attribute A one needs
  - the **Entropy** of a classification
  - the **Expectation Entropy** of the attribute A

Entropy („disorder“)

- Entropy is a measure of *(im)purity* of *a collection S of examples*.
- The higher the homogeneity of the information content, the lower the entropy
  - Let there be two classes + (positive) and – (negative).
  - Let p be the frequency of positive elements in S and n the frequency of negative elements in S
  - *The more **equal** p and n, the **higher** is the entropy*
    *the more **unequal** p and n, the **smaller** is the entropy*

high entropy 

low entropy 

Calculation of Entropy in Information Theory

- The defining expression for entropy in the theory of information was established by Claude E. Shannon in 1948
- It is of the form:

$$H = -\sum_i p_i \log_b p_i,$$

where

p_i is the probability of the message m_i

b is the base of the logarithm used (common values of b are 2, e and 10)

$\log_2(0)$ cannot be calculated; in the case of $p_i = 0$ for some *i*, the value of the corresponding summand $0 \log_b(0)$ is taken to be 0, which is consistent with the limit: $\lim_{p \to 0+} p \log(p) = 0$

Calculation of the Entropy for Binary Classification

- Assume a data set S with elements belonging to two classes C_1 and C_2
- The entropy is calculated by

$$Entropy\ (S) = -\ p_1 * log_2\ (p_1) -\ p_2 * log_2\ (p_2)$$

$\mathbf{p_i}$ relative frequencies of elements belonging to classes C_1 and C_2

$$p_1 = \frac{|C_1|}{|S|} \qquad p_2 = \frac{|C_2|}{|S|}$$

where

$|C_1|$ frequency of elements belonging to class C_1

$|C_2|$ frequency of elements belonging to class C_2

$\mathbf{|S| = |C_1| + |C_2|}$ is the number of all elements

Entropy Calculation for different Distributions

- The more different $|C_1|$ and $|C_2|$, the lower is the entropy

$\|C1\|$	$\|C2\|$	*p1*	*ld(p1)*	*p2*	*ld(p2)*	*Entropy(S)*
7	7	0.5	-1	0.5	-1	1
6	8	0.43	-1.22	0.57	-0.81	0.99
5	9	0.36	-1.49	0.64	-0.64	0.94
4	10	0.29	-1.81	0.71	-0.49	0.86
3	11	0.21	-2.22	0.79	-0.35	0.75
2	12	0.14	-2.81	0.86	-0.22	0.59
1	13	0.07	-3.81	0.93	-0.11	0.37

ld = $\log_2$ (logarithmus dualis)
$\log_2(0)$ cannot be calculated, but if a class is empty, i.e. $|C_1| = p_1 = 0$ or $|C_2| = p_2 = 0$ no classification is necessary. In this case $p_i * \log_2(p_i)$ is taken to be 0

Information Gain

- The information gain for an attribute A is the expected reduction in entropy caused be partitioning the example according to the attribute A
- The information gain is calculated by subtracting the expectation entropy of the subtrees created by A from the current entropy

$$GAIN\ (S, A) = Entropy\ (S) - EE(A)$$

Expected Entropy

- The **Expected Entropy EE_A** for an attribute A is the weighted average of the entropies of the subtrees created by the values v_i of A
- Let A be an attribute with m possible values v_1, ..., v_i, ... v_m
  - *Values(A)* is the set of all possible values for attribute A
  - S_v is the subset of S for which attribute A has value *v*
- The attribute A divides the elements into m partitions (subtrees)
- Entropy(S_v) is the entropy of the subtree for which the attribute A has value v

$$EE(A) := \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Formula for the Information Gain

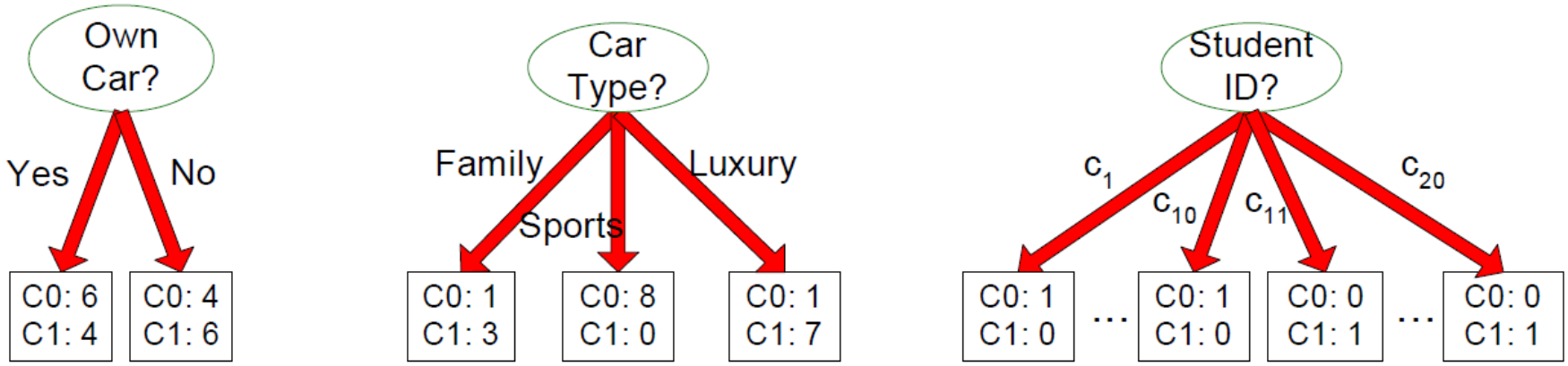
- The information gain for an attribute A is the expected reduction in entropy caused be partitioning the example according to the attribute A

$$GAIN(S, A) = Entropy(S) - \left(\sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \right)$$

Exercise

Entropy (S) = 1

Before Splitting: 10 records of class 0,
10 records of class 1



- Which test condition is the best?
- Does it make sense?

Thanks to Nadeem Qaisar Mehmood

ID3: Information Gain for Attribute Selection

- The goal of learning is to create a tree with minimal entropy
- ID3 uses the Information Gain to select the test attribute

On each level of the tree select the attribute with the highest information gain

- The recursive calculation of the attributes stops when either
  - all partitions contain only positive or only negative elements (i.e. entropy is 0) or
  - a user-defined threshold is achieved

Illustrative Example for ID3 Induction

An Illustrative Example (1)

The dependent variable „Tennis“ determines if the weather is good for tennis („Yes“) or not („No“).

Element	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Tennis*
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

An Illustrative Example (2): Entropy of the Decision Tree

$$\text{Entropy}(S) = - 9/14 * \log_2 (9/14) - 5/14 * \log_2 (5/14)$$
$$= 0{,}94$$

Element	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Tennis*
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

positive frequency (Yes)
negative frequency (No)

An Illustrative Example (3): Selection of the topmost Node

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

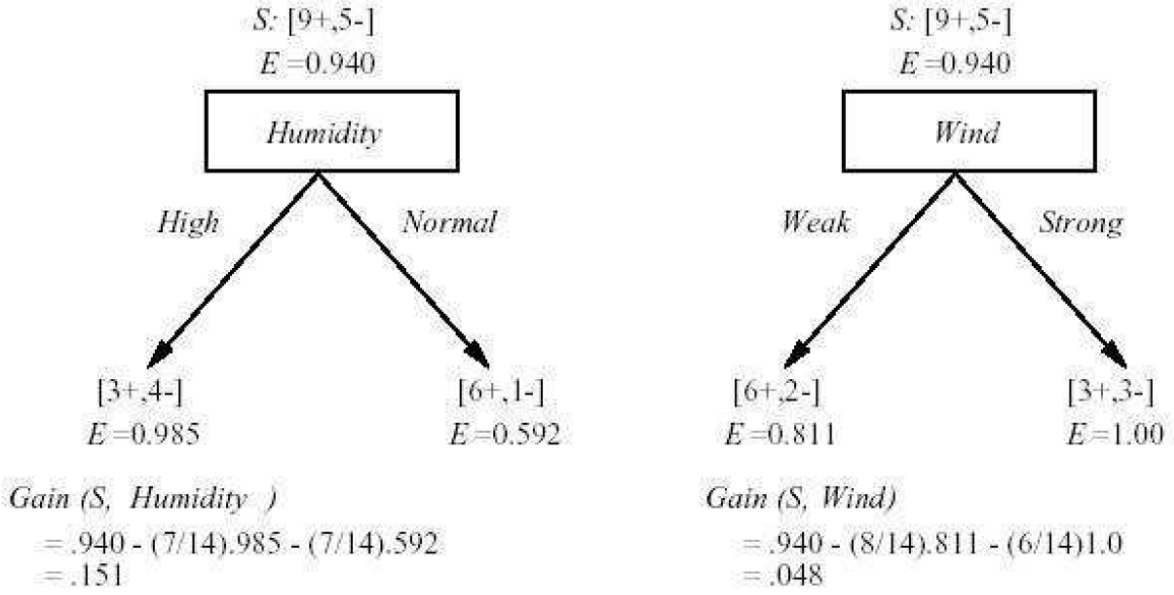
- In order to determine the attribute that should be tested first in the tree, the information gain for each attribute (*Outlook* , *Temperature, Humidity* and *Wind*) is determined.
  - Gain(S,Outlook) = **0.246**
  - Gain(S,Humidity) = **0.151**
  - Gain(S,Wind) = **0.048**
  - Gain(S,Temperature) = **0.029**
- Since *Outlook* attribute provides the best prediction, it is selected as the decision attribute for the root node.

An Illustrative Example (4): Computation of Information Gain

- The computation of Information Gain for Outlook:

$$GAIN(S,Outlook) = Entropy(S) - EE(Outlook)$$
$$= 0.94 - 0.694 = \mathbf{0.246}$$

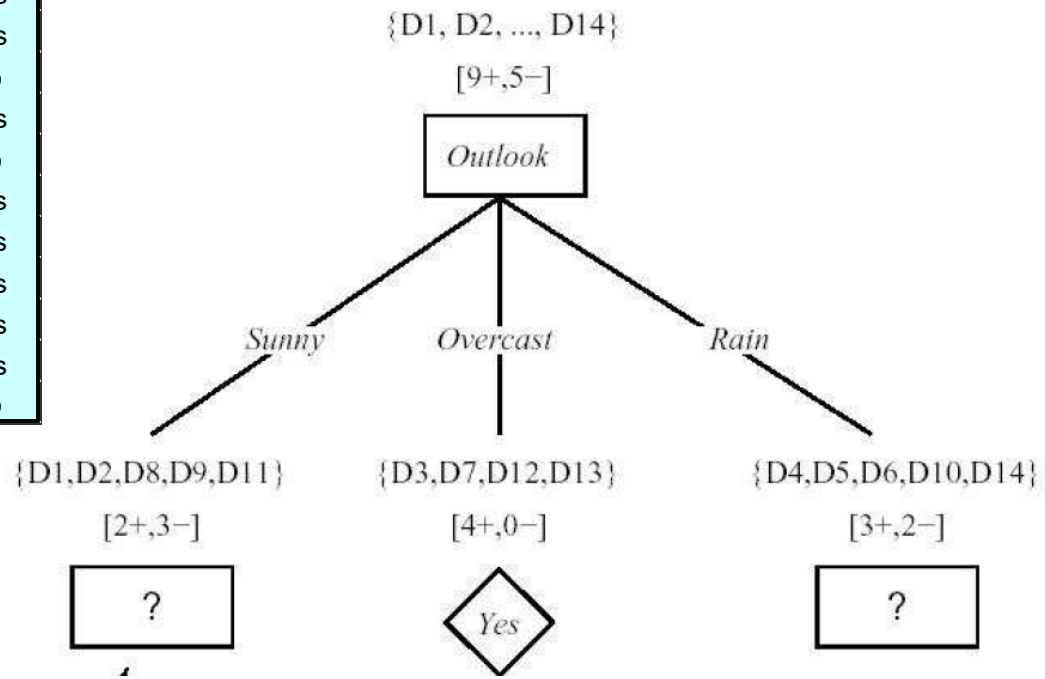
- The computation of information gain for *Humidity* and *Wind:*



An Illustrative Example (5): Resulting Subtree

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- The partially learned decision tree resulting from the first step of ID3:



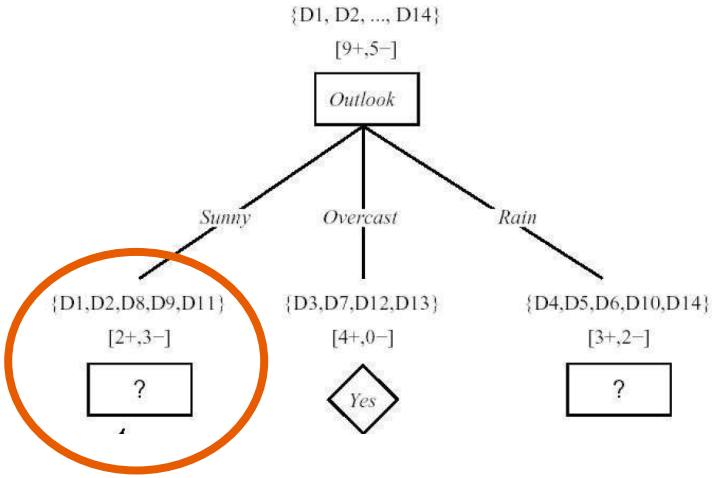
An Illustrative Example (6): Entropie of a Subtree

The subtree with root Sunny:

Entropy(Sunny) = $- 2 / 5 \log_2 (2/5) - 3 / 5 \log_2 (3/5)$

= 0,970

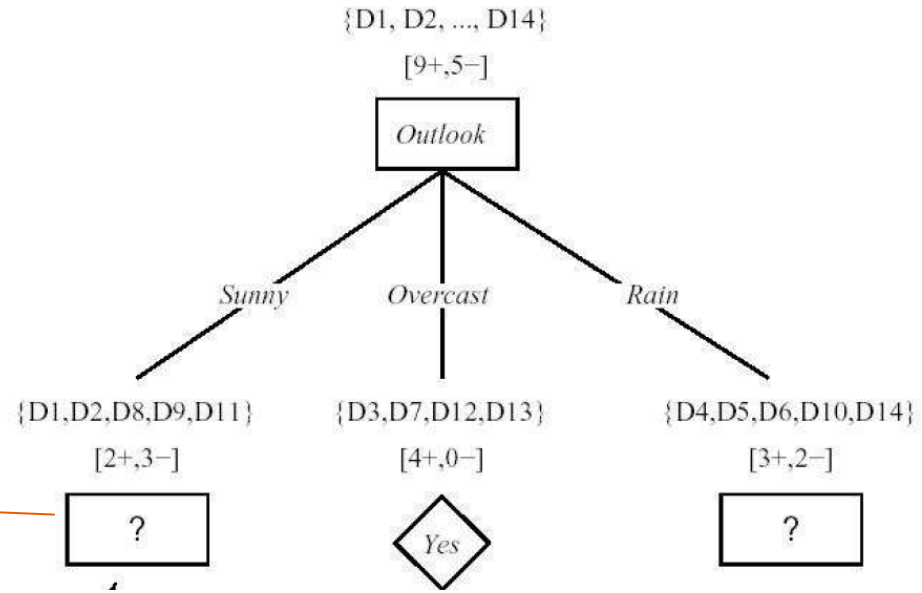
Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



The more **up** in the decision tree, the higher the entropy of the subtree

An Illustrative Example (7): Selectiong Next Attribute

Which attribute should be tested here?



$$S_{sunny} = \{D1, D2, D8, D9, D11\}$$

$$Gain\ (S_{sunny}, Humidity) = .970 - (3/5)\ 0.0 - (2/5)\ 0.0 = .970$$

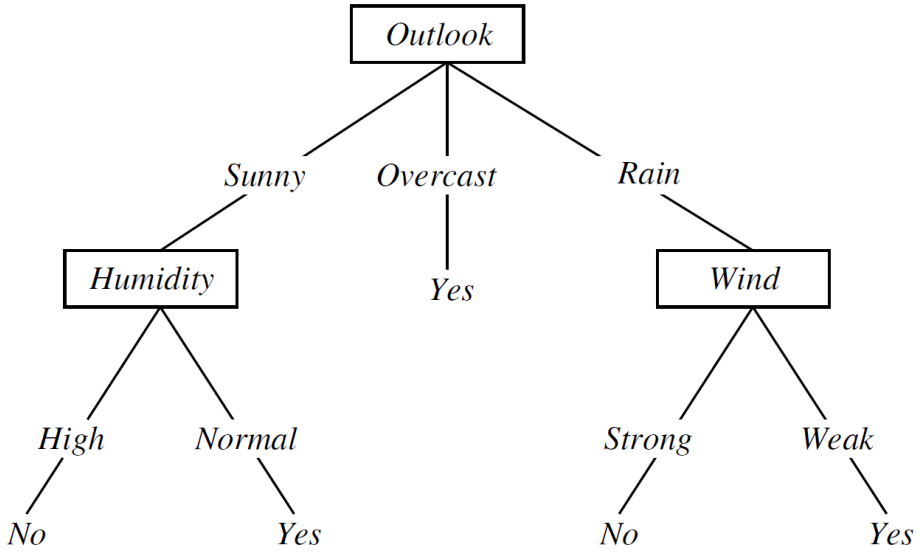
$$Gain\ (S_{sunny}, Temperature) = .970 - (2/5)\ 0.0 - (2/5)\ 1.0 - (1/5)\ 0.0 = .570$$

$$Gain\ (S_{sunny}, Wind) = .970 - (2/5)\ 1.0 - (3/5)\ .918 = .019$$

An Illustrative Example (8): The Resulting Decision Tree

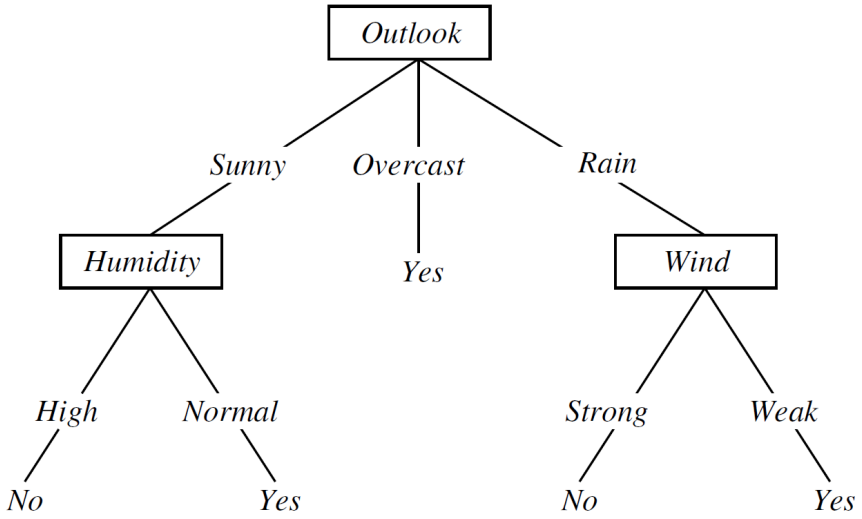
The dependent variable „Tennis“ determines if the weather is good for tennis („Yes“) or not („No“).

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



The result of the induction algorithms classifies the data with only three of the four attributes into the classes „Yes“ and „No“.

An Illustrative Example (9): Decision Tree represented as Decision Table



Playing Tennis				
	Outlook	Humidity	Wind	Tennis
	Sunny,Overcast, Rain	*High, Normal*	*Strong,Weak*	*Yes, No*
1	Sunny	High		No
2	Sunny	Normal		Yes
3	Overcast			Yes
4	Rain		Strong	No
5	Rain		Weak	Yes

Enhancements and Optimization

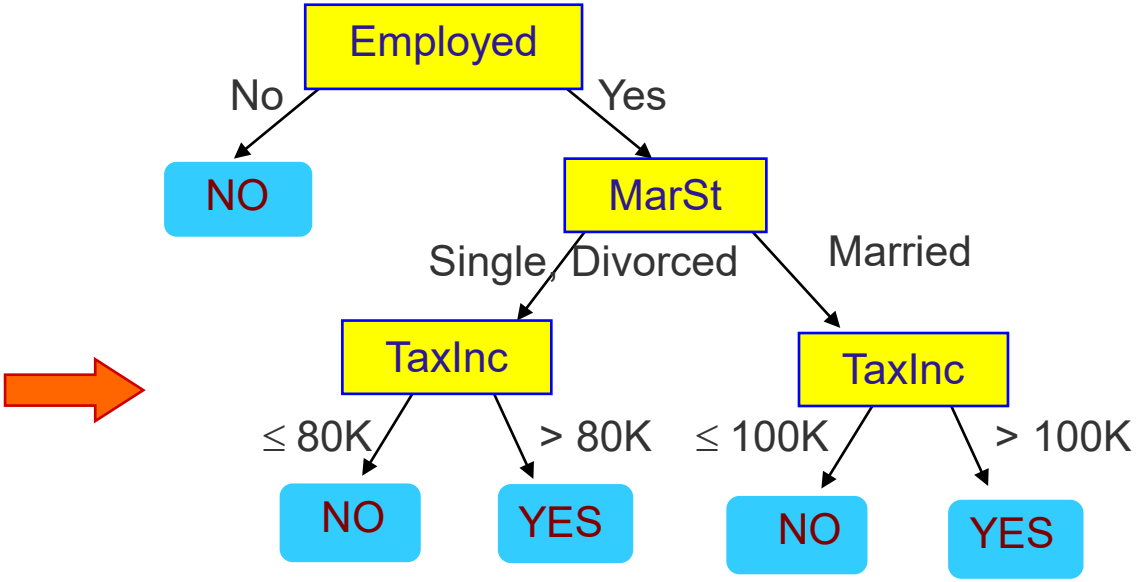
How to specify Attribute Test Conditions

Specification of the test condition depends on

- attribute types
  - Nominal
  - Ordinal
  - Continuous
- number of ways to split
  - 2-way split
  - Multi-way split

Splitting of Nominal and Numerical Data

Tid	Employed (categorical)	Marital Status (categorical)	Taxable Income (continuous)	accept (class)
1	No	Single	125K	No
2	Yes	Married	160K	Yes
3	Yes	Single	70K	No
4	No	Married	120K	No
5	Yes	Divorced	95K	Yes
6	Yes	Married	60K	No
7	No	Divorced	220K	No
8	Yes	Single	85K	Yes
9	Yes	Married	95K	No
10	Yes	Single	90K	Yes

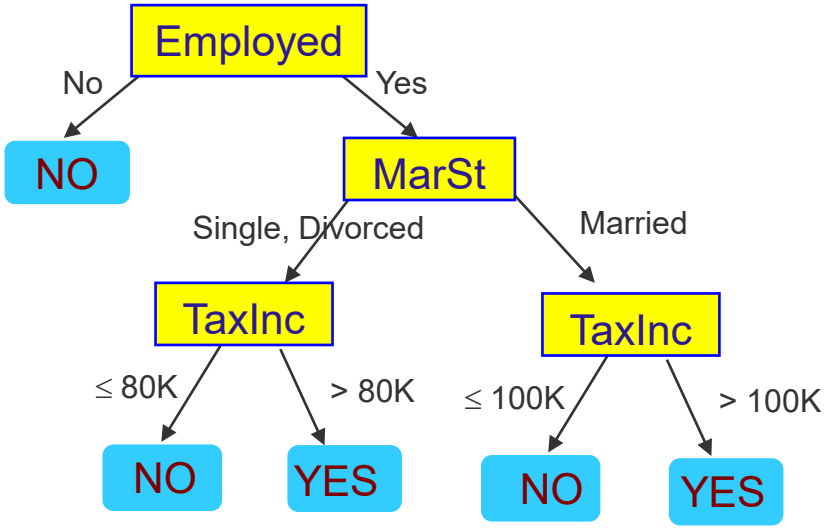


The model uses intervals for splitting numerical data

Splitting of Nominal and Numerical Data

Tid	Employed (categorical)	Marital Status (categorical)	Taxable Income (continuous)	accept (class)
1	No	Single	125K	No
2	Yes	Married	160K	Yes
3	Yes	Single	70K	No
4	No	Married	120K	No
5	Yes	Divorced	95K	Yes
6	Yes	Married	60K	No
7	No	Divorced	220K	No
8	Yes	Single	85K	Yes
9	Yes	Married	95K	No
10	Yes	Single	90K	Yes

Training Data



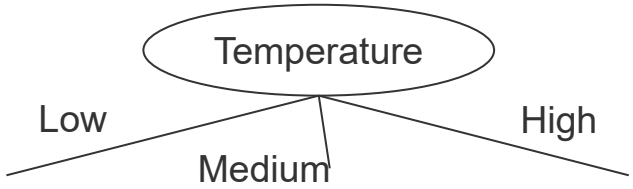
Model: Decision Tree

Credit Worthiness	Employed	Marital Status	Taxable Income	Accept
	Yes, No	*Single, Divorced, Married*	*Integer*	*Yes, No*
1	No			No
2	Yes	Single	> 80K	Yes
3	Yes	Divorced	> 80K	Yes
4	Yes	Single	≤ 80K	No
5	Yes	Divorced	≤ 80K	No
6	Yes	Married	> 100K	Yes
7	Yes	Married	≤ 100K	No

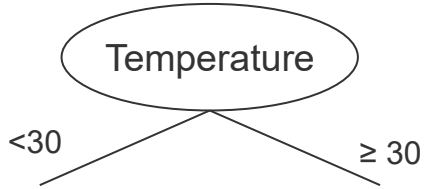
Model: Decision Table

Splitting for Continuous and Numerical Attributes

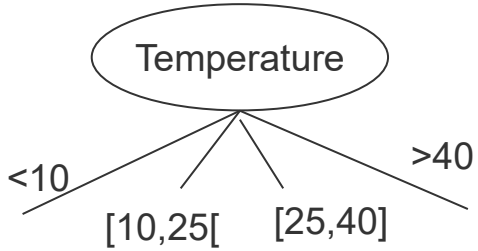
- Different ways of handling
  - Discretization to form an ordinal categorical attribute



  - Binary Decision: (A < v) or (A ≥ v)



  - Multi-way Split: Intervals



considering all possible splits and finding the best cut can be computing intensive

Preference for Short Trees

- Preference for short trees over larger trees, and for those with high information gain attributes near the root
- **Occam's Razor:** Prefer the simplest hypothesis that fits the data.
- Arguments in favor:
  - a short hypothesis that fits data is unlikely to be a coincidence – compared to long hypothesis
- Arguments opposed:
  - There are many ways to define small sets of hypotheses

Overfitting

- When there is ***noise in the data***, or when the number of training ***examples is too small***, the rule set (hypothesis) ***overfits*** the training examples!!
- Consider error of hypothesis *h* over
  - training data: *errortrain*(h)
  - entire distribution D of data: *errorD*(h)
- Hypothesis *h* ***overfits*** training data if there is an alternative hypothesis h*0* such that
  - *errortrain*(*h*) < *errortrain*(*h0*)
  - *errorD*(*h*) > *errorD*(*h0*)

Avoiding Overfitting by Pruning

- The classification quality of a tree can be improved by cutting weak branches
- Reduced error pruning
  - remove the subtree rooted at that node,
  - make it a leaf,
  - assign it the most common classification of the training examples afiliated with that node.
- To test accuracy, the data are separated in training set and validation set. Do until further pruning is harmful:
  - Evaluate impact on *validation* set of pruning each possible node
  - Greedily remove the one that most improves *validation* set accuracy

Pruning

These figures show the structure of a decision tree before and after pruning