DECISION TREES

Introduction

Classification

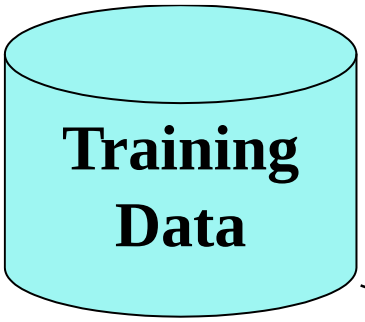
- **The goal of classification is to build a model based on the training data**
  - **Supervised learning technique**
- **Constructed by analyzing database samples**
- **Once such a predictive model is built**
  - **It can be used to predict the class of the objects of test cases**
- **Two sets of data in classification task**
  - **Training data, used to build the classification model**
  - **Test data, for which classes are not known**

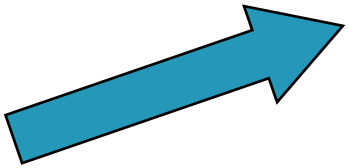
Classification—A Two-Step Process

- **Model construction: describing a set of predetermined classes**
  - **Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute**
  - **The set of tuples used for model construction: training set**
  - **The model is represented as classification rules, decision trees, or mathematical formulae**

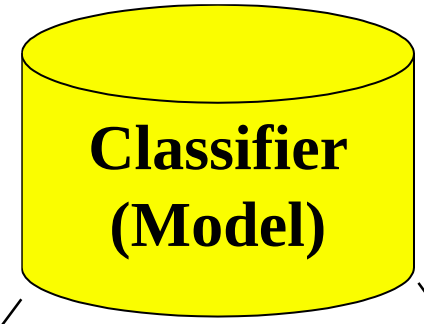
- **Model usage: for classifying future or unknown objects**
  - **Estimate accuracy of the model**
    - **The known label of test sample is compared with the classified result from the model**
    - **Accuracy rate is the percentage of test set samples that are correctly classified by the model**
    - **Test set is independent of training set, otherwise over-fitting will occur**

Classification Process (1): Model Construction





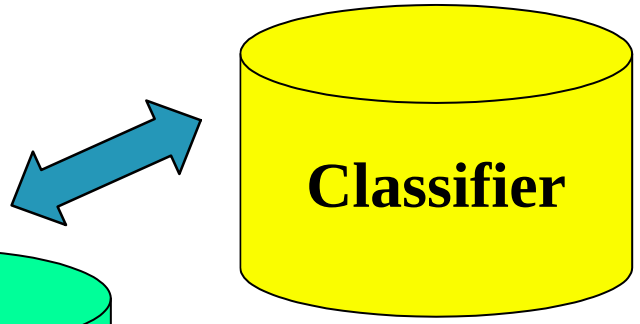
Classification
Algorithms

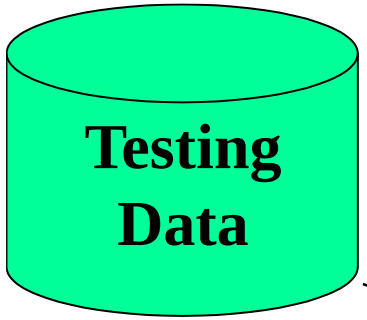


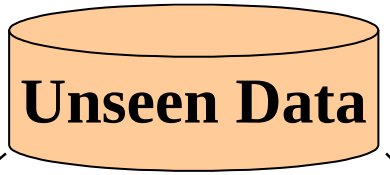
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Classification Process (2): Use the Model in Prediction







NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4)

Tenured?

Yes

Classification

- **Types of classifiers**
  - **Comprehensible classifiers (Rule based classifiers)**
    - **Decision tree, Ripper, CN2 etc.**
  - **Non-comprehensible classifiers (Statistical or mathematical classifiers)**
    - **SVM, Naïve bayes, NN etc.**
- **The aims of the research in classification field is**
  - **Find highly accurate learning models**
  - **Easy to understand (Comprehensible)**
  - **Efficient when dealing with large databases**

Evaluating Classification Methods

- **Predictive accuracy**
- **Speed and scalability**
  - **time to construct the model**
  - **time to use the model**
- **Robustness**
  - **handling noise and missing values**
- **Scalability**
  - **efficiency in disk-resident databases**
- **Interpretability**
  - **understanding and insight provided by the model**

DECISION TREES

Introduction

It is a method that induces concepts from examples (inductive learning)

Most widely used & practical learning method

The learning is *supervised*: i.e. the classes or categories of the data instances are known

It represents concepts as *decision trees* (which can be rewritten as if-then rules)

DECISION TREES

Introduction

The target function can be Boolean or discrete valued

DECISION TREES

Decision Tree Representation

1. Each node corresponds to an attribute

2. Each branch corresponds to an attribute value

3. Each leaf node assigns a classification

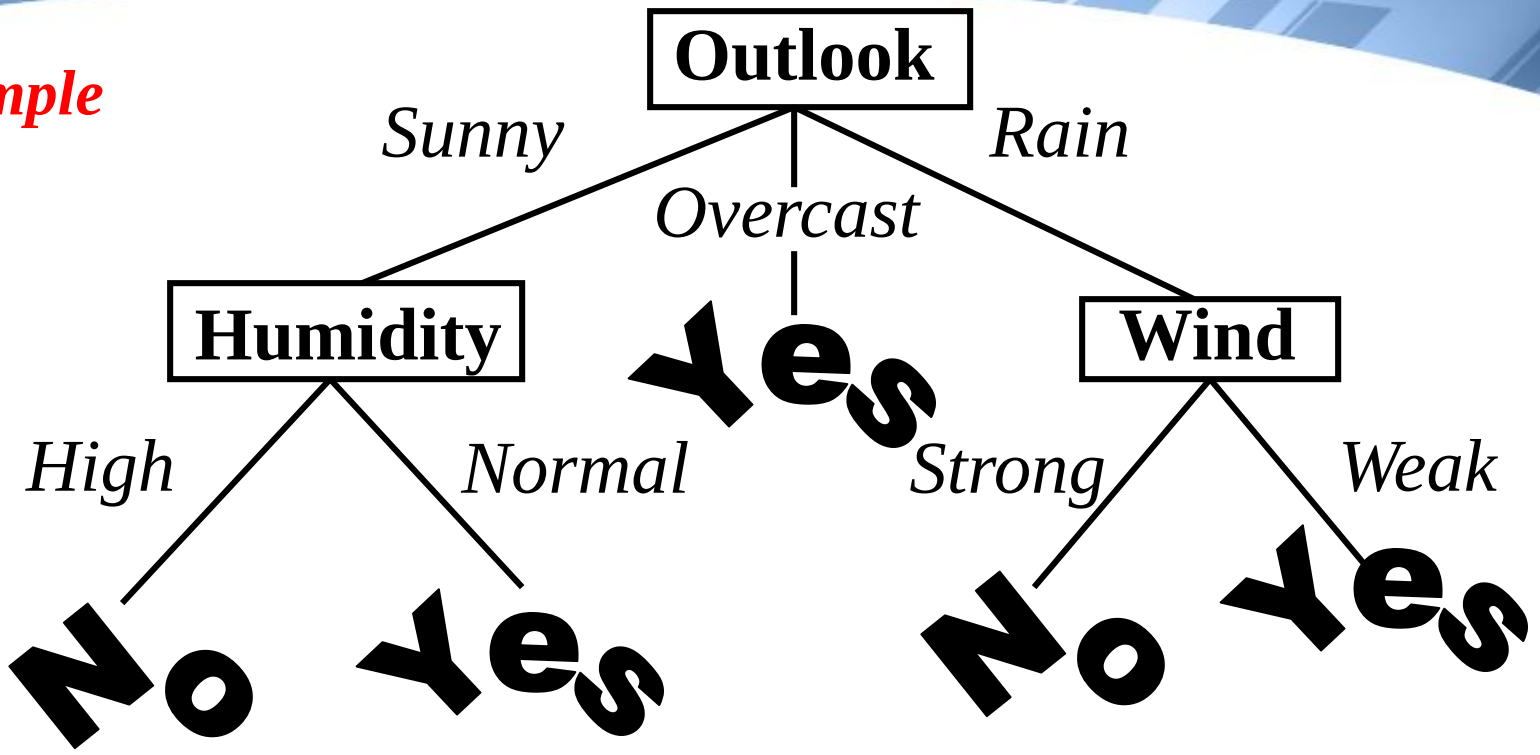
DECISION TREES

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

DECISION TREES

Example



A Decision Tree for the concept *PlayTennis*
An unknown observation is classified by testing its attributes and reaching a leaf node

DECISION TREES

Decision Tree Representation

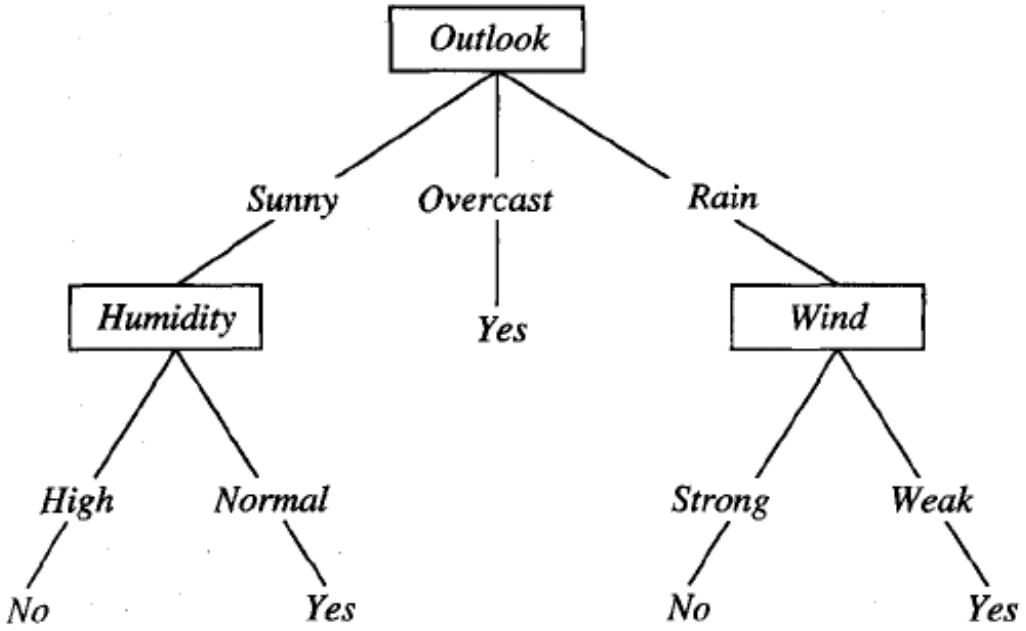
Decision trees represent a disjunction of conjunctions of constraints on the attribute values of instances

Each path from the tree *root* to a *leaf* corresponds to a conjunction of attribute tests (one rule for classification)

The tree itself corresponds to a disjunction (OR) of these conjunctions (AND) (set of rules for classification)

DECISION TREES

Decision Tree Representation



$$(Outlook = Sunny \wedge Humidity = Normal)$$
$$\vee \quad (Outlook = Overcast)$$
$$\vee \quad (Outlook = Rain \wedge Wind = Weak)$$

DECISION TREES

Basic Decision Tree Learning Algorithm

DECISION TREES

Basic Decision Tree Learning Algorithm

Most algorithms for growing decision trees are variants of a basic algorithm

An example of this core algorithm is the ID3 algorithm developed by Quinlan (1986)

It employs a top-down, greedy search through the space of possible decision trees

DECISION TREES

Basic Decision Tree Learning Algorithm

First of all we *select* the best attribute to be tested at the root of the tree

For making this selection each attribute is evaluated using a statistical test to determine how well it alone classifies the training examples

The selection process is then repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree

DECISION TREES

Basic Decision Tree Learning Algorithm

This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices

DECISION TREES

Which Attribute is the Best Classifier?

DECISION TREES

Which Attribute is the Best Classifier?

The central choice in the ID3 algorithm is selecting which attribute to test at each node in the tree

We would like to select the attribute which is most useful for classifying examples

For this we need a good quantitative measure

For this purpose a statistical property, called *information gain* is used

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

In order to define information gain precisely, we begin by defining entropy

Entropy is a measure commonly used in information theory.

Entropy characterizes the impurity of an arbitrary collection of examples

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

This formula is called Entropy H

$$\mathbf{H(X)} = -\sum_{j=1}^{m} p_j \log_2 p_j$$

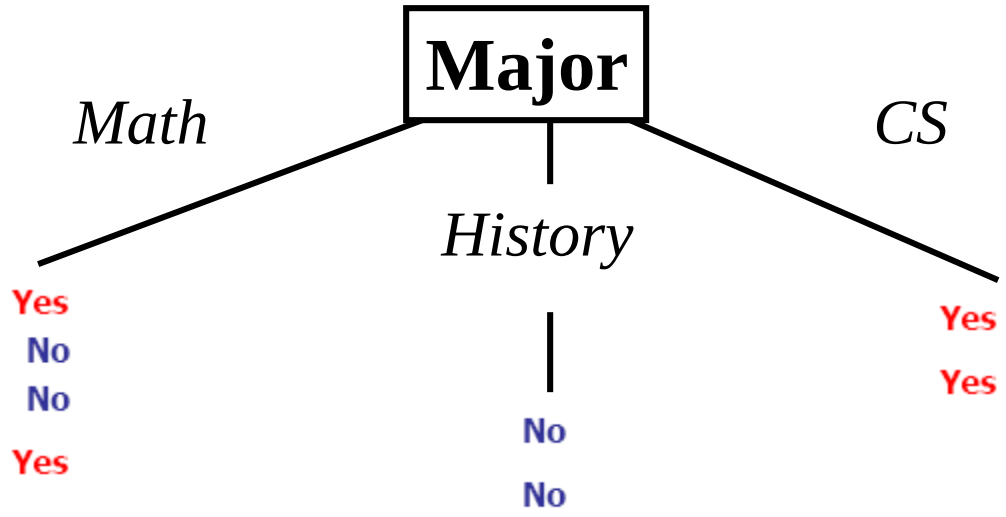
High Entropy means that the examples have almost equal probability of occurrence (and therefore not easily predictable)

Low Entropy means easy predictability

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

Suppose we are trying to predict output Y & we have input X (College Major = v)



X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

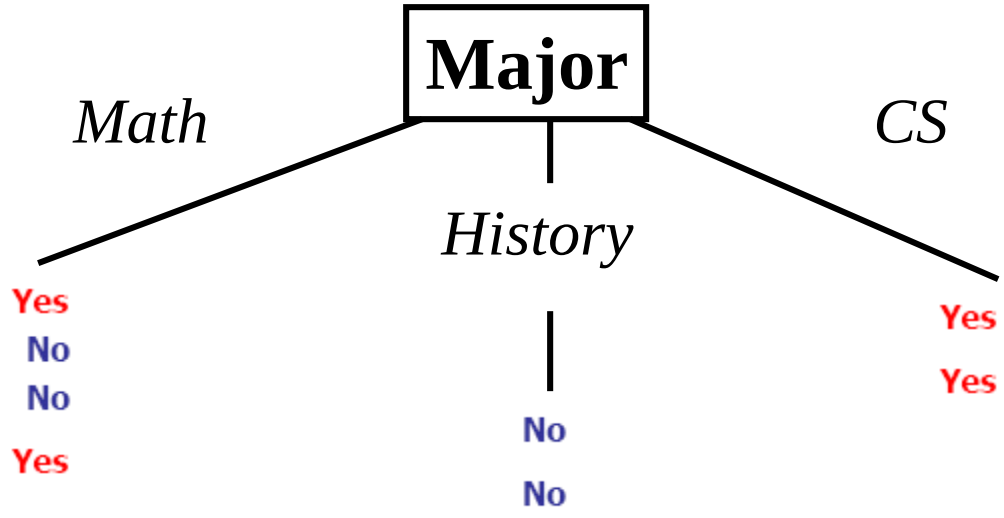
DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

We have H(Y) = 1.0

Conditional Entropy H(Y | X = v)

The Entropy of Y among only those records in which X = v



X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

DECISION TREES

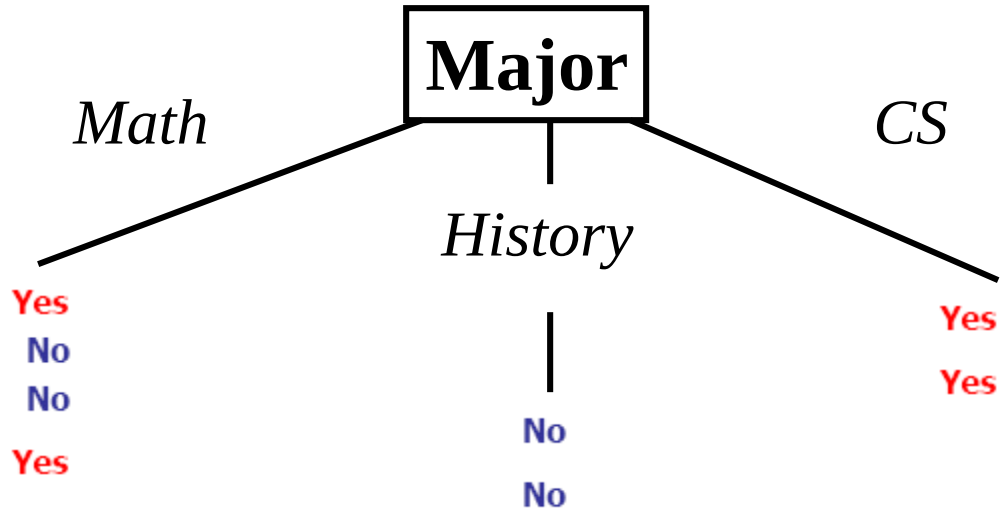
Which Attribute is the Best Classifier?: Information Gain

Conditional Entropy of Y

H(Y | X = Math) = 1.0

H(Y | X = History) = 0

H(Y | X = CS) = 0



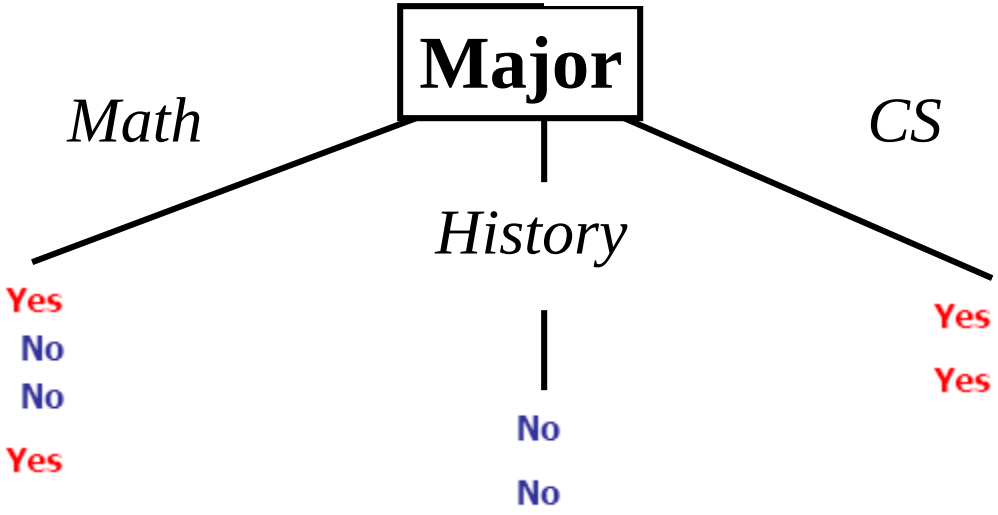
X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

Average Conditional Entropy of Y

$$\mathbf{H(Y \mid X)} = \Sigma_j \, Prob(X=v_j) \, H(Y \mid X = v_j)$$



v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$$

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

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

Let's investigate the attribute *Wind*

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

The collection of examples has 9 positive values and 5 negative ones

$$Entropy(S) = 0.940$$

Eight (6 positive and 2 negative ones) of these examples have the attribute value *Wind* = *Weak*

Six (3 positive and 3 negative ones) of these examples have the attribute value *Wind* = *Strong*

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

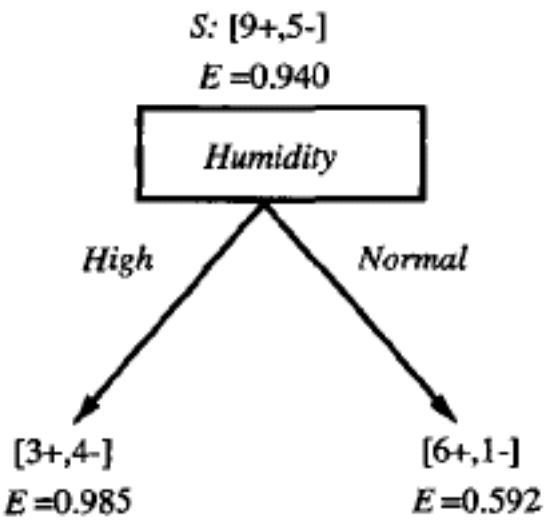
The information gain obtained by separating the examples according to the attribute *Wind* is calculated as:

$$
\begin{aligned}
Gain(S, Wind) &= Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S) - (8/14) Entropy(S_{Weak}) \\
&\quad - (6/14) Entropy(S_{Strong}) \\
&= 0.940 - (8/14)0.811 - (6/14)1.00 \\
&= 0.048
\end{aligned}
$$

DECISION TREES

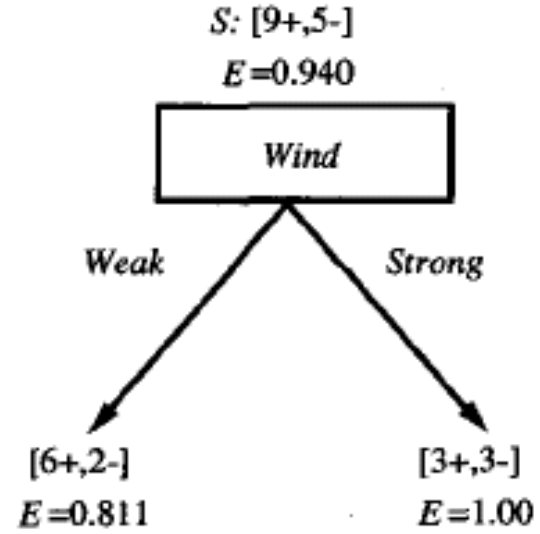
Which Attribute is the Best Classifier?: Information Gain

We calculate the Info Gain for each attribute and select the attribute having the highest Info Gain



Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151



Gain (S, Wind)

= .940 - (8/14).811 - (6/14)1.0
= .048

DECISION TREES

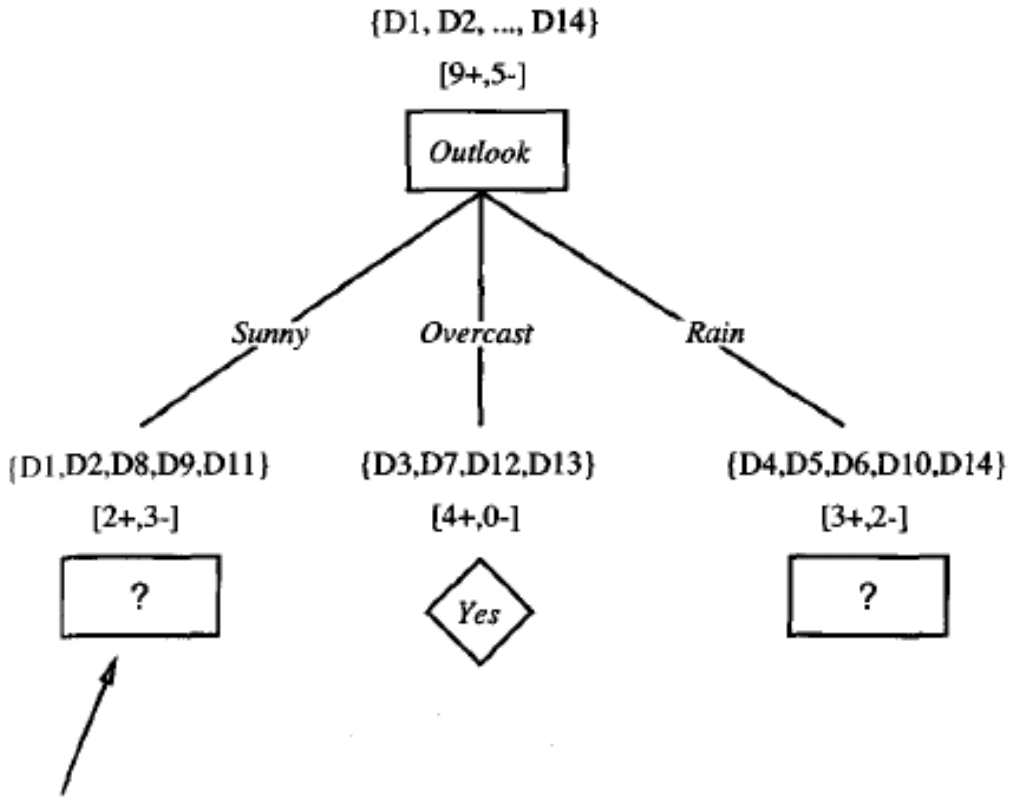
Example

Which attribute should be selected as the first test?

$$Gain(S, Outlook) = 0.246$$
$$Gain(S, Humidity) = 0.151$$
$$Gain(S, Wind) = 0.048$$
$$Gain(S, Temperature) = 0.029$$

“Outlook” provides the most information

DECISION TREES



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

DECISION TREES

Example

The process of selecting a new attribute is now repeated for each (non-terminal) descendant node, this time using only training examples associated with that node

Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree

DECISION TREES

Example

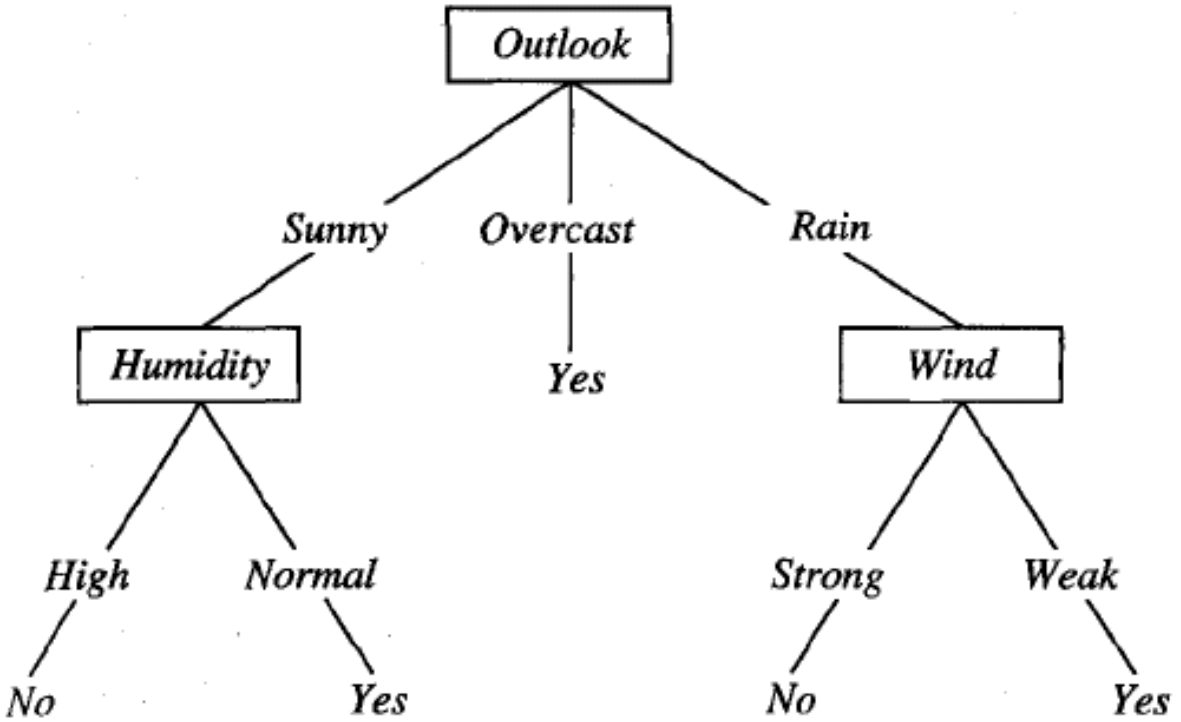
This process continues for each new leaf node until either:

1. **Every attribute has already been included along this path through the tree**

2. **The training examples associated with a leaf node have zero entropy**

DECISION TREES

Example



DECISION TREES

From Decision Trees to Rules

Next Step: Make rules from the decision tree

After making the decision tree, we trace each path from the root node to leaf node, recording the test outcomes as *antecedents* and the leaf node classification as the *consequent*

For our example we have:

If the Outlook is Sunny and the Humidity is High then No
If the Outlook is Sunny and the Humidity is Normal then Yes
...

DECISION TREES

Example

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

	Grades	***Hardworking***	***Intelligent***	***Unlucky***
1.	**Good**	**Yes**	**Yes**	**No**
2.	**Bad**	**No**	**Yes**	**Yes**
3.	**Bad**	**Yes**	**No**	**Yes**
4.	**Good**	**Yes**	**Yes**	**No**
5.	**Good**	**Yes**	**Yes**	**No**
6.	**Bad**	**No**	**Yes**	**No**
7.	**Bad**	**Yes**	**No**	**No**

DECISION TREES

Grades	Hardworking	Intelligent	Unlucky
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Show the steps of decision tree induction algorithm for the data

The concepts to be learnt are Grades (Grades are class labels)

If two or more attributes have similar Information Gain, pick any of them.

Since the best entropy is zero, hence if the average entropy of an attribute is zero, there is no need to check other candidate attributes.

DECISION TREES

Grades	Hardworking	Intelligent	Unlucky
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

The entropy of the data:

There are 3 samples of Good Grades and 4 samples of Bad Grades.

Hence the entropy the data is

$$- (3/7) \log_2 (3/7) - (4/7) \log_2 (4/7)$$
$$= 0.524 + 0.461$$
$$= 0.985$$

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

First we calculate which of the three attributes will be best for the root node.

There are three attributes:

Hardworking, Intelligent, Unlucky

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Attribute "Hardworking"

Branch Hardworking = Yes (5 samples fulfill this condition)
Branch Hardworking = No (2 samples fulfill this condition)

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Attribute "Hardworking"

Branch Hardworking = Yes (5 samples fulfill this condition)

There are 3 samples of Good Grades and
2 samples of Bad Grades.

Entropy of this branch =

$$-(3/5)\log_2(3/5) - (2/5)\log_2(2/5)\}$$
$$= 0.442 + 0.53$$
$$= 0.972$$

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Attribute "Hardworking"

Branch Hardworking = No (2 samples fulfill this condition)

Both of these 2 samples are of Bad Grades

Entropy of this branch =

$$- (0/2) \log_2 (0/2) - (2/2) \log_2 (2/2) = 0$$

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Average Entropy of Attribute "Hardworking" =

5/7 * 0.972 + 2/7 * 0 = 0.694

Information Gain of Attribute "Hardworking" =

0.985 – 0.694 = 0.291

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Similarly we calculate the average entropy and information gain of the attributes "Intelligent" and "Unlucky"

We need not calculate these values fully because both of the attributes have

- **two branches**
- **one branch has 5 & the other one has 2 samples**
- **The branch with 5 samples has 3 samples of a certain class label and 2 samples of another class label**
- **The branch with 2 samples has homogenous class labels**

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Anyway after calculation (or after inference due to symmetry) we find that the information gain of all three attributes is similar

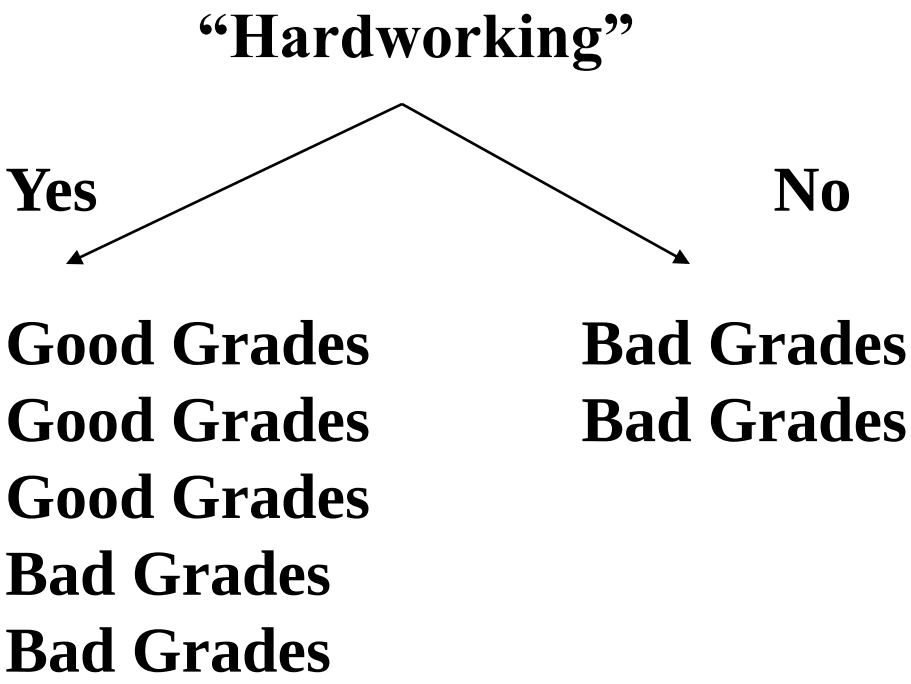
Hence we can place any attribute at the root node

Arbitrarily we choose “Hardworking”

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

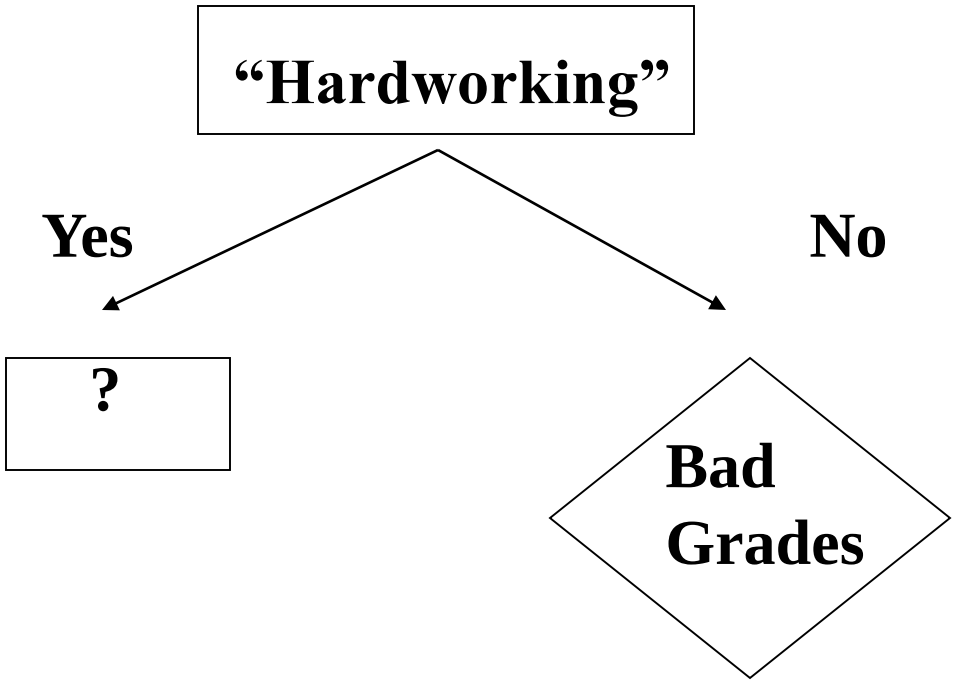


Since 2nd branch has samples of homogenous class labels hence we do not expand it further

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example



DECISION TREES

Grades	Hardworking	Intelligent	Unlucky
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

The branch "Hardworking" = Yes has to be grown

We test the two remaining attributes "Intelligent" and "Unlucky"

First we note the Entropy of the branch "Hardworking" = Yes for the purpose of Information Gain

There are a total of 5 samples: 3 samples of Good Grades and 2 samples of Bad Grades

Entropy is

$$-(3/5)\log_2(3/5) - (2/5)\log_2(2/5)$$

$$= 0.442 + 0.53 = 0.972$$

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Attribute "Intelligent"

Branch Intelligent = Yes (3 samples fulfill this condition)

All the 3 samples are of Good Grades.

Hence Entropy = 0

Branch Intelligent = No (2 samples fulfill this condition)

Both of these 2 samples are of Bad Grades.

Hence Entropy = 0

DECISION TREES

Grades	*Hardworking*	*Intelligent*	*Unlucky*
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example

Attribute "Intelligent"

Average Entropy of Attribute "Intelligent"

= 3/5 * 0 + 2/5 * 0 = 0

Information Gain of Attribute "Intelligent"

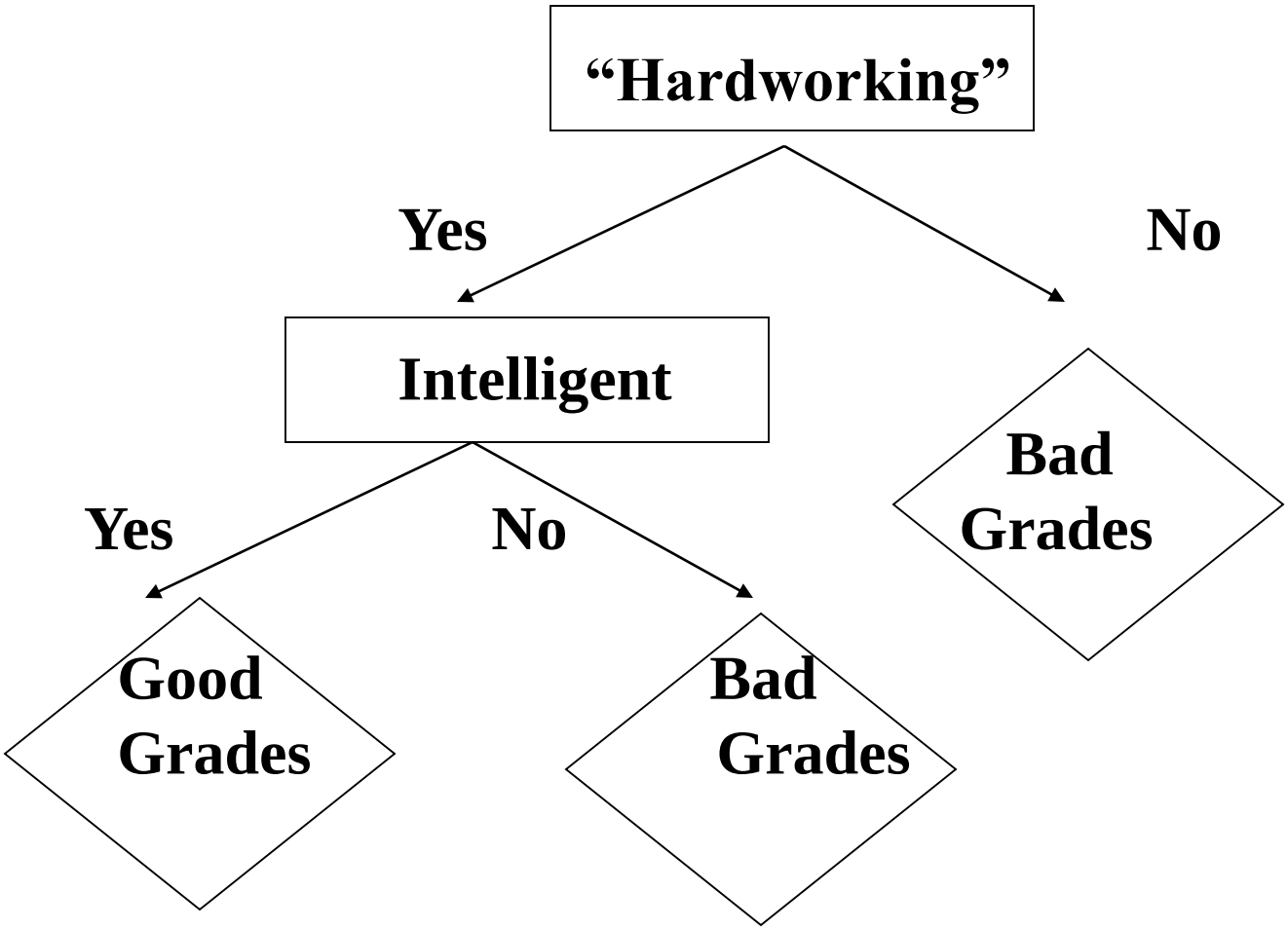
= 0.972 – 0 = 0.972

Since we cannot have better results, therefore there is no need to test the other attribute "Unlucky".

DECISION TREES

Grades	Hardworking	Intelligent	Unlucky
Good	Yes	Yes	No
Bad	No	Yes	Yes
Bad	Yes	No	Yes
Good	Yes	Yes	No
Good	Yes	Yes	No
Bad	No	Yes	No
Bad	Yes	No	No

Example



DECISION TREES

Reference

Sections 3.1 – 3.4 of T. Mitchell

Section 4.3 of Witten & Frank