MACHINE INTELLIGENCE

DECISION TREE -ID3 ALGORITHM

K.S.Srinivas
Department of Computer Science and Engineering

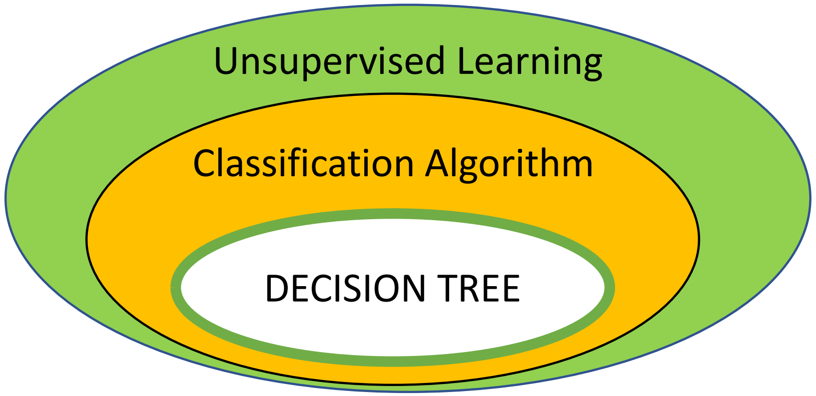
MACHINE INTELLIGENCE

DECISION TREE -ID3 ALGORITHM

Srinivas K S.
Associate Professor, Department of Computer Science

What is a Decision Tree algorithm?

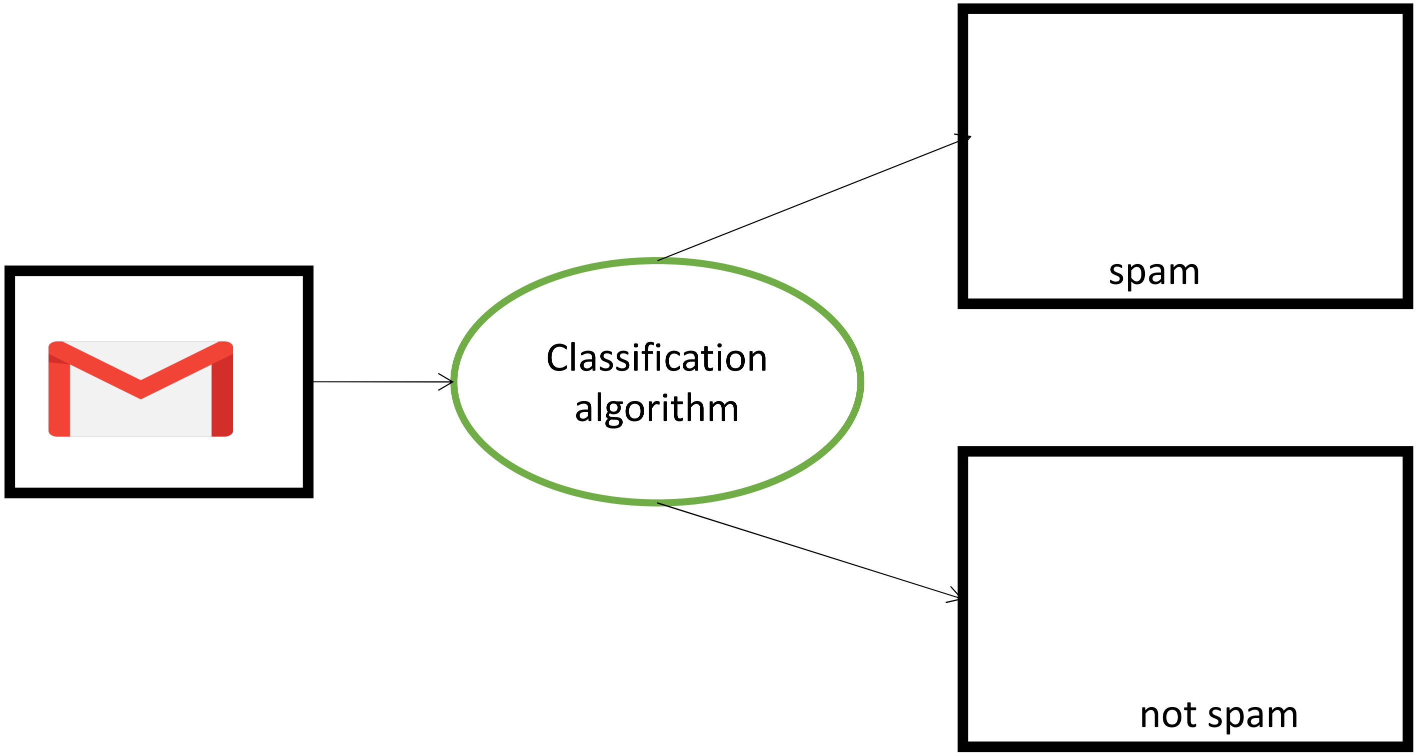
- A type of classification algorithm
- Comes under unsupervised learning technique



MACHINE INTELLIGENCE
What is classification algorithm?

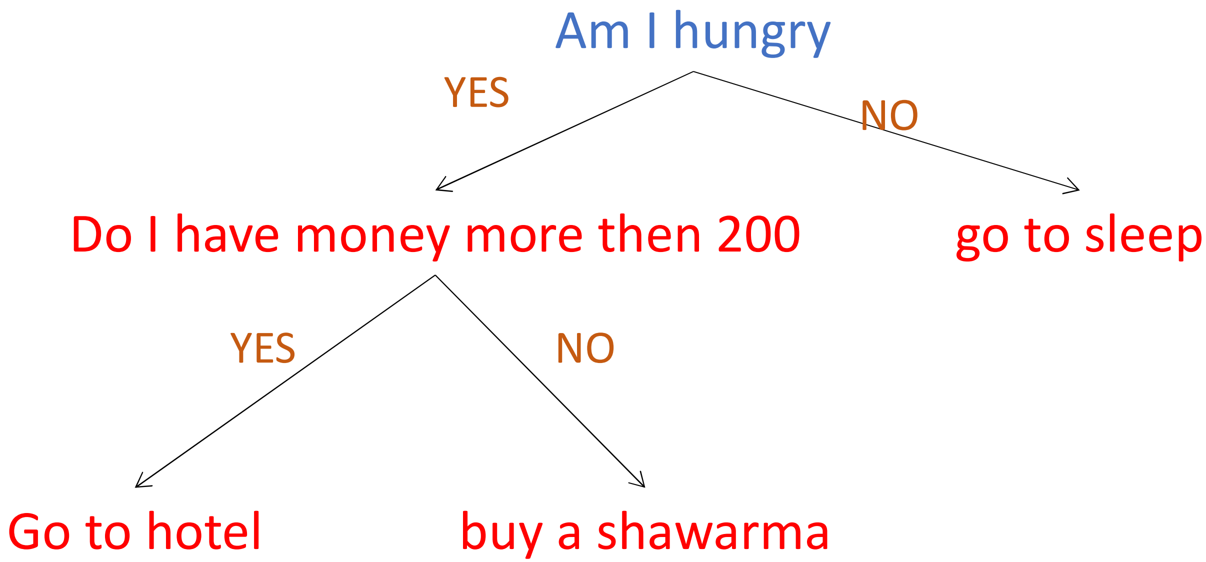
"Classification is the process of dividing the data sets into different categories or groups by adding label"

example: classification of emails as spam or not spam based on certain conditions



What is a Decision Tree?

- Graphical representation of all the possible solutions to a decision
- Decisions are based on some conditions
- Decision made can be easily explained

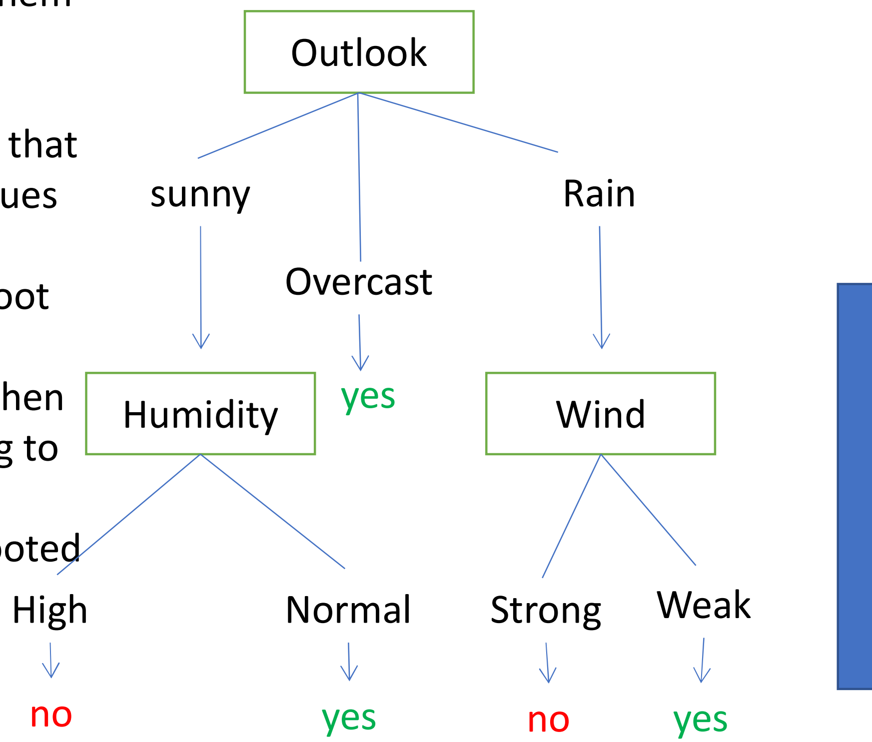


Decision Tree Learning definition-Tom Mitchell

"Decision tree learning is a method for approximating discrete-valued target function,in which the learned function is represented by a decision tree."

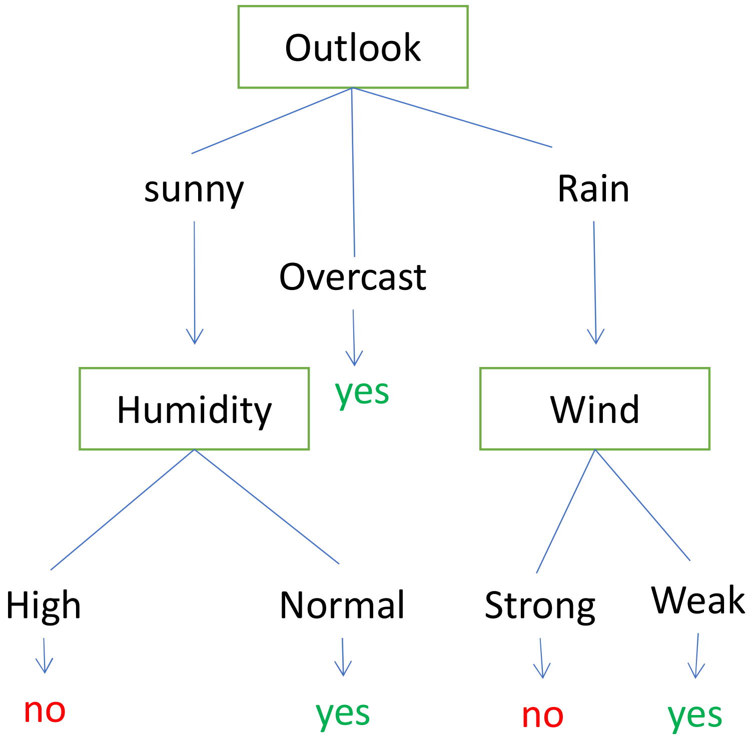
- Decision trees classify instances by sorting them down the tree from the root to some leaf node ,which provides classification of the instances and each branch descending from that node corresponds to one of the possible values of this attribute.
- An instances is classified by starting at the root node of the tree
- testing the attribute specified by this node then moving down the tree branch corresponding to value of the attribute.
- This process is then repeated for sub tree rooted at the new node

Decision tree for concept to play Tennis



Decision Function

- Based on this decision tree lets try to think of the function that our machine may have learn't



f(outlook,humidity,wind)= {

1 **if** (outlook=sunny and humidity=Normal)
or
(outlook=overcast)
or
(outlook=Rain and Wind=weak)

0 for any other inputs }

Appropriate Problem for Decision Tree Learning

- **Instances are Represented by attribute-value pairs**
  example: Temperature and their values
- **The target function has discrete output values**
  example the previous example of concept of playing tennis
- **Disjunctive descriptions may be required**
  As noted above, decision trees naturally represent disjunctive expressions.
- **The training data may contain errors**
  Decision tree learning methods are robust to errors, both errors in classifications of the training examples and errors in the attribute values that describe these examples.
- **The training data may contain missing attribute values**
  Decision tree methods can be used even when some training examples have unknown values

ID3 Algorithm

- In this course we will look at the basic ID3 algorithm for learning decision trees
- Later we will examine the hypothesis space search performed by this learning algorithm.
- We will then head forward to characterize the inductive bias of this ID3 algorithm
- At last we will see the problem of over fitting the training data , also check strategies to deal with it.

ID3 Approach

- Our basic algorithm ,ID3 learns decision trees by constructing them top-down, beginning with question .

  "which attribute should be tested at the root of the tree?"
- The simple answer **statistical test**
- We evaluate each instance attribute using statistical test to determine how well it alone classifies the training example.
- The best attribute is selected and used as the test at the root node of the tree.
- A descendant of the root node is then created for each possible value of this attribute, and the training examples are sorted to the appropriate descendant node.
- The entire 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. This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices.
- We will define a statistical property, called information gain, that measures how well a given attribute separates the training examples according to their target classification.

Entropy

- In order to define information gain precisely, we begin by defining a measure commonly used in information theory, called entropy
- characterizes the (im)purity of an arbitrary collection of examples.
- Given a collection S, containing positive **(p)** and negative **(n)** examples of some target concept, the entropy of S relative to this Boolean classification is

$$\text{Entropy(S)} = \boxed{\frac{-\mathrm{p}}{\mathrm{p}+n}\log_2\left(\frac{p}{p+n}\right)} - \boxed{\frac{n}{n+p}\log_2\left(\frac{n}{n+p}\right)}$$

uncertainty due to positive examples in data set

uncertainty due to positive examples in data set

- More generally, if the target attribute can take on c different values, then the entropy of S relative to this c-wise classification is defined as

$$\text{Entropy(S)} = \sum_{\mathrm{i}=1}^{\mathrm{c}} -\mathrm{p}_{\mathrm{i}} \log_2(p_i)$$

where p_{i} is the proportion of S belonging to class i.

Entropy Calculation

consider the following dataset of accepting a job

first we will calculate the entropy of entire data set S

$$\text{Entropy(S)} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{n+p}\log_2\left(\frac{n}{n+p}\right)$$

Now let us calculate total number of positive and negative points

number of positive samples p=3

number of negative samples n = 4

salary	Location	job acceptance
Tier1	MUM	YES
Tier 2	BLR	YES
Tier 1	BLR	NO
Tier 1	HYD	NO
Tier 2	MUM	YES
Tier 1	HYD	NO
Tier 1	HYD	NO

$$\text{Entropy(S)} = \boxed{\frac{-3}{3+4}\log_2\left(\frac{3}{3+4}\right)} - \boxed{\frac{4}{4+3}\log_2\left(\frac{4}{4+3}\right)}$$

$= \boxed{-0.428 \times (-1.222)} - \boxed{0.5714 \times -0.8}$

=0.98

Entropy Calculation

$$\text{Entropy}(S) = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{n+p}\log_2\left(\frac{n}{n+p}\right)$$

now lets take the attribute salary to split our data
we will first calculate the entropy of salary=Tier1

number of positive points = 1

number of negative points = 7

$$\text{Entropy}(\text{Salary} = \text{Tier1}) = \boxed{\frac{-1}{1+7}\log_2\left(\frac{1}{1+7}\right)} - \boxed{\frac{7}{7+1}\log_2\left(\frac{7}{1+7}\right)}$$

$= \boxed{0.375} - \boxed{(-0.16856)}$

=0.543

salary	Location	job acceptance
Tier1	MUM	YES
Tier 1	BLR	NO
Tier 1	HYD	NO
Tier 1	HYD	NO
Tier 1	HYD	NO
Tier 1	HYD	NO
Tier 1	HYD	NO

salary	Location	job acceptance
Tier 2	BLR	YES
Tier 2	MUM	YES

salary	entropy
Tier1	0.543

MACHINE INTELLIGENCE

Entropy Calculation

$$\text{Entropy}(S) = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{n+p}\log_2\left(\frac{n}{n+p}\right)$$

we will calculate entropy of salary=Tier2

number of positive points = 2

number of negative points = 0

$$\text{Entropy}(\text{Salary}=\text{Tier2}) = \boxed{\frac{-2}{2+0}\log_2\left(\frac{2}{2+0}\right)} - \boxed{\frac{0}{2+0}\log_2\left(\frac{0}{2+0}\right)}$$

$= \boxed{0} - \boxed{0}$

$=0$

Note: ENTROPY =0 when all samples are of one class
ENTROPY=1 when all class have equal samples

salary	Location	job acceptance
Tier1	MUM	YES
Tier 1	BLR	NO
Tier 1	HYD	NO
Tier 1	HYD	NO
Tier 1	HYD	NO
Tier 1	HYD	NO
Tier 1	HYD	NO

salary	Location	job acceptance
Tier 2	BLR	YES
Tier 2	MUM	YES

salary	entropy
Tier1	0.543
Tier2	0

Average Information

The statistical term Average Information of a attribute is given by

$$\mathrm{I(Attribute)} = \sum \frac{\mathrm{p_i} + n_i}{p+n} Entropy(A)$$

Lets us understand this by using the previous calculations we did

salary	entropy
Tier1	0.543
Tier2	0

now,

$$\mathrm{I(SALARY)} = \frac{\mathrm{p_{tier1}} + n_{tier1}}{p+n} Entropy(salary = tier1) +$$

$$\frac{p_{tier2} + n_{tier2}}{p+n} Entropy(salary = tier2)$$

$$\mathrm{I(SALARY)} \quad = \frac{1+7}{3+7} \times 0.543 + \frac{2+0}{3+7} \times 0$$

=0.4344

Information Gain

- Given entropy as a measure of the impurity in a collection of training examples, we can now define a measure of the effectiveness of an attribute in classifying the training data
- The measure we will use, called **information gain**
- Is simply the expected reduction in entropy caused by partitioning the examples according to this attribute
- More precisely Information gain G(S,A) of an attribute A relative to collection of example S is defined by

G(S, A)=ENTROPY(S)-I(A)

that is differences of entropy of the collection of example S and information gain of the attribute A

Information Gain- Calculation

from our previous calculation we have the following data

Entropy(S)	0.98
I(salary)	0.4344

let us calculate the Information gain G(S, Salary)

G(S, A)=ENTROPY(S)-I(A)

G(S,SALARY)=ENTROPY(S)-I(SALARY)

G(S,SALARY)=0.98-0.4344
=0.5456

The ID3 Algorithm

Steps to create a decision tree using the ID3 algorithm

1. COMPUTE THE **ENTROPY** FOR DATA-SET **ENTROPY(S)**
2. FOR EVERY ATTRIBUTE
   - CALCULATE ENTROPY FOR ALL OTHER VAUES **ENTROPY(A)**
   - TAKE **AVERAGE INFORMATION ENTROPY** FOR THE CURRENT ATTRIBUTE
   - CLACULATE **GAIN** FOR THE CURRENT ATTRIBUTE
3. PICK THE **HIGHEST GAIN ATTRIBUTE**
4. **REPEAT** UNTIL WE GET THE TREE WE DESIRED

The ID3 Algorithm- problem

Let us now use all the previous knowledge and create a decision tree of the following data set

the play tennis is the target attribute here

Outlook	Temp	Humidity	Windy	Play tennis
Sunny	High	High	Weak	No
Sunny	High	High	Strong	No
Overcast	High	High	Weak	Yes
Rainy	Medium	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Medium	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Medium	Normal	Weak	Yes
Sunny	Medium	Normal	Strong	Yes
Overcast	Medium	High	Strong	Yes
Overcast	High	Normal	Weak	Yes
Rainy	Medium	High	Strong	No

Lets create decision tree for this by following the steps

step 1: COMPUTE THE **ENTROPY** FOR DATA-SET

ENTROPY(S)

$$\text{Entropy(S)} = \frac{-\text{p}}{\text{p}+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{n+p}\log_2\left(\frac{n}{n+p}\right)$$

number of positive points = 9

number of negative points = 5

$$\text{Entropy(S)} = \frac{-9}{9+5}\log_2\left(\frac{9}{9+5}\right) - \frac{5}{5+9}\log_2\left(\frac{5}{5+9}\right)$$

=0.94

MACHINE INTELLIGENCE

The ID3 Algorithm- problem

2. FOR EVERY ATTRIBUTE
   - CALCULATE ENTROPY FOR ALL OTHER VAUES **ENTROPY(A)**
   - TAKE **AVERAGE INFORMATION ENTROPY** FOR THE CURRENT ATTRIBUTE
   - CLACULATE **GAIN** FOR THE CURRENT ATTRIBUTE

we will first check outlook attribute and create sub table for outlook =sunny,outlook =rainy, and outlook =overcast

for outlook=sunny

number of positive points = 2

number of negative points = 3

$$\text{Entropy(outlook} = \text{sunny)} = \frac{-2}{2+3}\log_2\left(\frac{2}{2+3}\right) - \frac{3}{2+3}\log_2\left(\frac{3}{2+3}\right)$$ =0.971

similarly for rainy and overcast we have the following results

$$\text{Entropy(outlook} = \text{overcast)} = -1\log_2(1) - 0\log_2(0) = 0$$

$$\text{Entropy(outlook} = \text{rainy)} = \frac{-3}{2+3}\log_2\left(\frac{3}{2+3}\right) - \frac{2}{2+3}\log_2\left(\frac{2}{2+3}\right) = 0.971$$

now we calculate Average information of the attribute outlook

$$I(outlook) = \frac{3+2}{9+5} * 0.971 + \frac{2+3}{9+5} * 0.971 + \frac{4+0}{9+5} * 0 = 0.693$$

finally we calculate Information gain for the attribute outlook

G(S,outlook)=0.94-0.693=0.247

Outlook	Play tennis
Sunny	No
Sunny	No
Sunny	No
Sunny	Yes
Sunny	Yes

property	value
Entropy(s)	0.94
G(outlook)	0.247

The ID3 Algorithm- problem

we will do the same procedures for other table and the obtain the following result

property	value
Entropy(s)	0.94
G(outlook)	0.247
G(temp)	0.029
G(humidity)	0.152
G(windy)	0.048

The ID3 Algorithm- problem

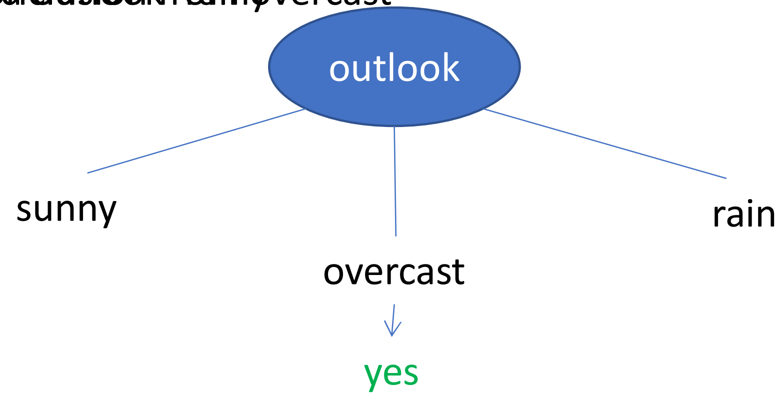
3. PICK THE **HIGHEST GAIN ATTRIBUTE**
4. **REPEAT** UNTIL WE GET THE TREE WE DESIRED

we will now create our tree using the 3rd and 4th step

we choose our root node as outlook and split according to its three values

remember entropy of overcast split is 0 which means it is pure hence we can come to a conclusion from overcast

outlook=sunny and outlook=rain



property	value
Entropy(s)	0.94
G(outlook)	0.247
G(temp)	0.029
G(humidity)	0.152
G(windy)	0.048

The ID3 Algorithm- problem

with outlook=sunny our data would look something like this

first we calculate entropy of the data set

p=2 n=3

$$\text{Entropy}(\text{S}_{\text{sunny}}) = \frac{-2}{2+3}\log_2\left(\frac{2}{2+3}\right) - \frac{3}{3+2}\log_2\left(\frac{3}{3+2}\right) = 0.97$$

next we need to consider each attribute and calculate its gain

Outlook	Temp	Humidity	Windy	Play tennis
Sunny	High	High	Weak	No
Sunny	High	High	Strong	No
Sunny	Medium	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Medium	Normal	Strong	Yes

The ID3 Algorithm- problem

considering temperature attribute

Temperature	p	n	Entropy
cool	1	0	0
high	0	2	0
medium	1	1	1

Outlook	Temp	Play tennis
Sunny	High	No
Sunny	High	No
Sunny	Medium	No
Sunny	Cool	Yes
Sunny	Medium	Yes

Average Information Entropy: I(Temp)=0.4

Gain : G(Temp)=0.571

The ID3 Algorithm- problem

considering humidity attribute

Temperature	p	n	Entropy
normal	2	0	0
high	0	3	0

Outlook	Humidity	Play tennis
Sunny	High	No
Sunny	High	No
Sunny	High	No
Sunny	Normal	Yes
Sunny	Normal	Yes

Average Information Entropy: I(Temp)=0
Gain : G(Temp)=0.971

The ID3 Algorithm- problem

considering windy attribute

Temperature	p	n	Entropy
strong	1	1	1
weak	1	2	0.918

Outlook	Windy	Play tennis
Sunny	Weak	No
Sunny	Strong	No
Sunny	Weak	No
Sunny	Weak	Yes
Sunny	Strong	Yes

Average Information Entropy: I(Temp)=0.951

Gain : G(Temp)=0.020

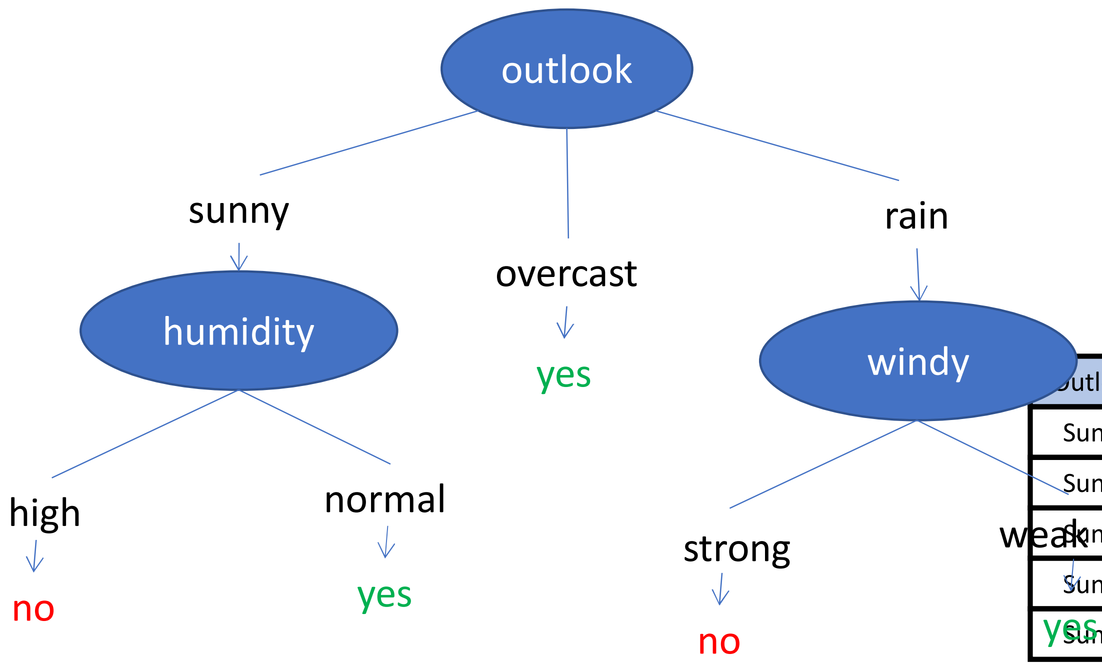
MACHINE INTELLIGENCE

The ID3 Algorithm- problem

after the previous calculations we have the following data for the data with outlook=sunny

[illegible]

property	value
Entropy(S_{sunny})	0.97
G(temp)	0.571
G(humidity)	0.971
G(windy)	0.02

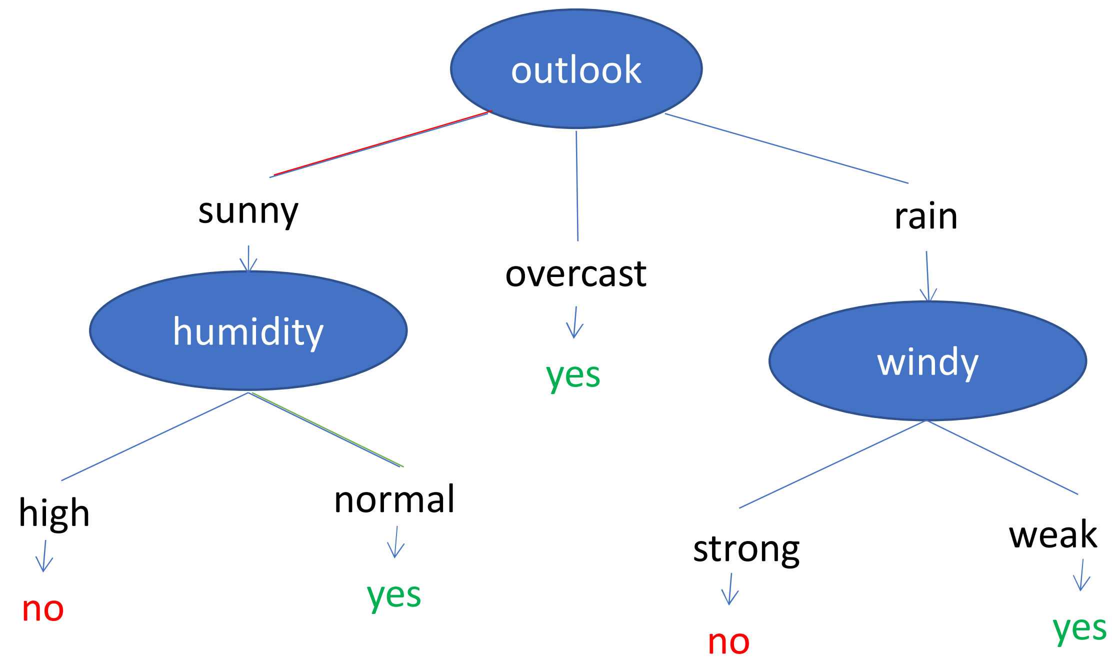


Outlook	Humidity	Play tennis
Sunny	High	No
Sunny	High	No
Sunny	High	No
Sunny	Normal	Yes
Sunny	Normal	Yes

Decision Tree analysis

Let us see how decision tree helps us predicting if a player will play tennis or not

on a given day let this be the report of the weather forecast
"the day would be sunny with normal humidity and weak wind"



PES UNIVERSITY ONLINE

THANK YOU

K.S.Srinivas
srinivasks@pes.edu
+91 80 2672 1983 Extn 701