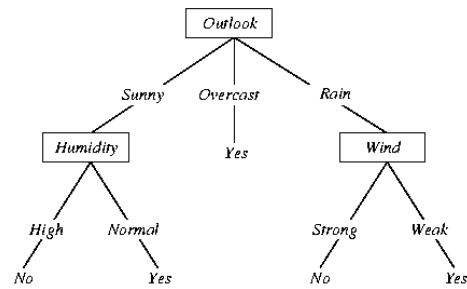


Decision Trees

Reading assignment: Textbook, Chapter 6

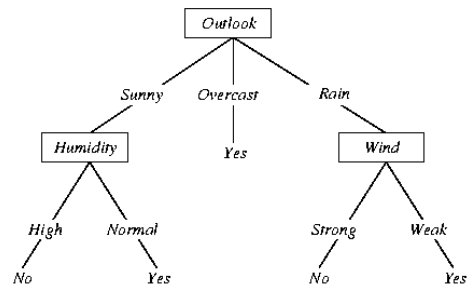
<u>Day</u>	<u>Outlook</u>	<u>Temp</u>	<u>Humidity</u>	<u>Wind</u>	<u>PlayTennis</u>
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	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



- Target concept: “Good days to play golf”
- Example:
<Outlook = Sunny, Temperature = Hot, Humidity = High, Wind = Strong>

Classification?

Decision Trees



Can represent this in logical form as disjunction of conjunctions:

- Would it be possible to use a “generate-and-test” strategy to find a correct decision tree?

- Why should we care about finding the simplest (i.e., smallest) correct decision tree?

Example: Detecting spam

```
From: Alibris <books@alibris.m0.net>  
Reply-to: books@alibris.m0.net  
To: mm@cse.ogi.edu  
Subject: **SPAM 06.70** Melanie, reminding you to save $10 at Alibris
```

HOLIDAY SPECIAL: SAVE UP TO \$10 ON YOUR PURCHASES
(order now and receive by Christmas)

With the holiday season rapidly approaching, we want to remind you of our most generous sale of the year. As a valued customer, we invite you to save up to \$10 off your Alibris purchases with three ways to save:

\$2 off your order of \$20 or more: GIFT2
\$5 off your order of \$50 or more: GIFT5
\$10 off your order of \$100 or more: GIFT10

Simply enter the coupon codes above* at checkout. But hurry, this limited time offer expires on December 16, 2003. Visit Alibris now and save!

Save money on shipping too! Now through December 9, 2003, every item listed on our site should be delivered to continental U.S. and Canadian addresses by December 24th via standard shipping (our lowest cost option) or get FREE shipping when you order \$49 of In Stock books. Don't delay, start your holiday shopping now.
<http://alibris.m0.net/m/S.asp?HB10950943733X2869462X274232X>

From: "Basil Lutz" <0eynsozueb@a-city.de>
 Reply-To: "Basil Lutz" <0eynsozueb@a-city.de>
 To: <mm@santafe.edu>, <bonabeau@santafe.edu>
 Subject: **SPAM 10.70** This tool will make your website more productive hukm

```
<html>
<head>
<title>hd36 8 ekj 009 920 2                </title>
<meta http-equiv=3D"Content-Type" content=3D"text/html; charset=3Diso-8859=
-1">
</head>

<body>
<p><font face=3D"Arial, Helvetica, sans-serif">Can your website answer que=
stions
  in real time 24 hours a day, 7 days a week? Our clients websites do and =
we're
  not talking about some stale FAQ sheet either. Add <a href=3D"http://www=
dreamscaper.co.mn@click.net-click.net.ph/click.php?id=3Ddrcomnm">live
  operator support</a> to your website today and dramatically increase you=
r revenues.</font></p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p>&nbsp;</p>
<p><a href=3D"http://www.dreamscaper.co.mn@click.net-click.net.ph/click.ph=
p?id=3Ddrcomnm">stop</a>
  sending me emails</p>
</body>
</html>
```

From: =?iso-8859-1?q?james=20ken?= <ja_ken2004@yahoo.fr>
 Subject: **SPAM 13.30** URGENT ASSISTANCE
 To: ja_ken2004@yahoo.fr

FROM: JAMES KEN.

ATTN:

Dear Respectful one,

I know this proposal letter may come to you as a surprise considering the fact that we have not had any formal acquaintance before .but all the same I would want you for the sake of God to give this an immediate attention in view of the fact that the security of our live and possession is at stake .

I am Mr JAMES KEN 28 years old from war ravaged SIERRA LEONE but presently domiciled in Abidjan Ivory coast with my sister JANET who is 18 years old .My father Mr KEN who before his untimely assassination by the rebels was the Director of SIERRA LEONE Diamond corporation (SLDC) .He was killed in our government residential house along side two of my other brothers ,two house maids and one government attached security guard fortunately for I, younger sister and mother ,we were on a week end visit to our home town As we got the news of the tragedy .We immediately managed to ran into neighbouring Ivory coast for refuge .But unfortunately .As Fate would have it ,we lost our dear mother (may soulrest in peace) as a result of what the Doctor called cardiac arrest .

As we were coming into this country ,we had some documents of a deposit of \$ 11 700 000 USD (eleven million seven hundred thousand USD) made by my late father in a security and trust company .According to my father, he intended to use this fund for his international business transaction after his tenure in office but was unfortunately murdered .We had located the security company where the money is deposited with the help of an attorney and established ownership .please right now ,with the bitter experiences we had in our country and the war still going on especially in diamond area which incidentally is where we hail from .coupled with the incessant political upheavals and hostilities in this country Ivory coast ,we desire seriously to leave here and live the rest of our life into a more peaceful and politically stable country like yours Hence this proposal and request .We therefore wish you can help us in the following regards :

- 1)To provide us with a good bank account to transfer the money into.
- 2)To help us invest the money into a lucrative business .
- 3)To assist my sister Janet get a college admission to further her education.

Please I know that , this letter may sound strange and incredible to you but the CNN and the BBC African bulletin normally have it as their major news features .Therefore for the sake of God and humanity give an immediate positive consideration and reply to me via our e-mail address. I will willingly agree to any suitable percentage of the money you will propose as your compensation for your assistance with regards to the above .please in view of our sensitive refugee status and as we are still conscious of our father 's enemies .I would like you to give this a highly confidential approach .

Best Regards .
JAMES KEN.

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
6.70 points, 4 required;
* 0.4 -- BODY: Offers a limited time offer
* 0.1 -- BODY: Free Offer
* 0.4 -- BODY: Stop with the offers, coupons, discounts etc!
* 0.1 -- BODY: HTML font color is red
* 0.1 -- BODY: Image tag with an ID code to identify you
* 2.8 -- BODY: Bayesian classifier says spam probability is 80 to 90%
[score: 0.8204]
* 0.8 -- BODY: HTML font color is green
* 0.3 -- BODY: FONT Size +2 and up or 3 and up
* 0.1 -- BODY: HTML font color not within safe 6x6x6 palette
* 0.1 -- BODY: HTML font color is blue
* 0.3 -- BODY: Message is 70% to 80% HTML
* 1.2 -- Date: is 6 to 12 hours after Received: date
---- End of SpamAssassin results
```

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
10.70 points, 4 required;
* 0.4 -- BODY: Message is 40% to 50% HTML
* 1.0 -- URI: URL contains username and (optional) password
* 0.8 -- URI: Uses a username in a URL
* 1.2 -- RBL: Received via a relay in dnsbl.njabl.org
[RBL check: found 78.199.241.24.dnsbl.njabl.org.,]
[type: 127.0.0.9]
* 4.3 -- RBL: Received via a relay in list.dsbl.org
[RBL check: found 78.199.241.24.list.dsbl.org.]
* 0.1 -- Message has X-MSMail-Priority, but no X-MimeOLE
* 0.1 -- Message only has text/html MIME parts
* 2.8 -- Forged mail pretending to be from MS Outlook IMO
---- End of SpamAssassin results
```

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
13.30 points, 4 required;
* 0.7 -- From: ends in numbers
* 1.4 -- Subject is indicative of a Nigerian spam
* 1.5 -- BODY: Nigerian scam key phrase (million dollars)
* 2.7 -- BODY: Contains urgent matter
* 2.8 -- BODY: Bayesian classifier says spam probability is 80 to 90%
      [score: 0.8584]
* 0.7 -- Subject is all capitals
* 0.7 -- From: contains an underline and numbers/letters
* 2.8 -- Message body has multiple indications of Nigerian spam
---- End of SpamAssassin results
```

Can we do such classifications with a decision tree?

What features should we use?

- [Spamassasin features](#)

Decision Tree Induction

- Goal is, given set of training examples, construct decision tree that will classify those training examples correctly (and, hopefully, generalize)
- Original idea of decision trees developed in 1960s by psychologists Hunt, Marin, and Stone, as model of human concept learning. (CLS = “Concept Learning System”)
- In 1970s, AI researcher Ross Quinlan used this idea for AI concept learning:
 - [ID3 \(“Itemized Dichotomizer 3”\)](#), 1979

The Basic Decision Tree Learning Algorithm (ID3)

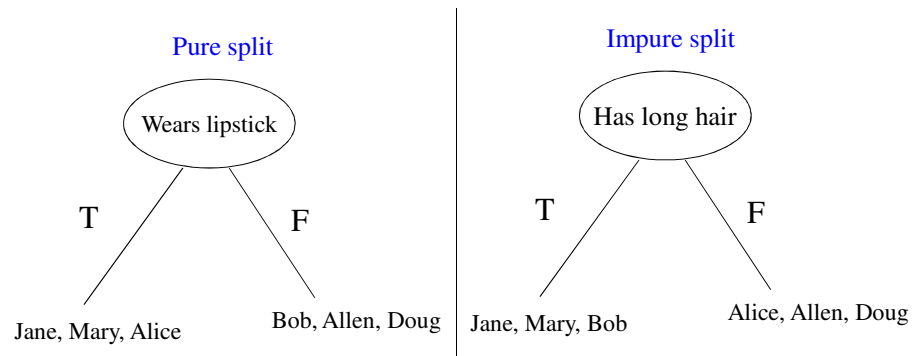
1. Determine which attribute is, by itself, the most useful one for distinguishing the two classes over all the training data. Put it at the root of the tree.
2. Create branches from the root node for each possible value of this attribute. Sort training examples to the appropriate value.
3. At each descendant node, determine which attribute is, by itself, the most useful one for distinguishing the two classes for the corresponding training data. Put that attribute at that node.
4. Go to 2, but for the current node.

Note: This is greedy search with no backtracking

How to determine which attribute is the best classifier for a set of training examples?

“Impurity” of a split

- Perfect (“pure”) split: all instances on a branch belong to same class.



For each node, we want to choose attribute that gives purest split.
But how to measure degree of impurity of a split ?

Entropy

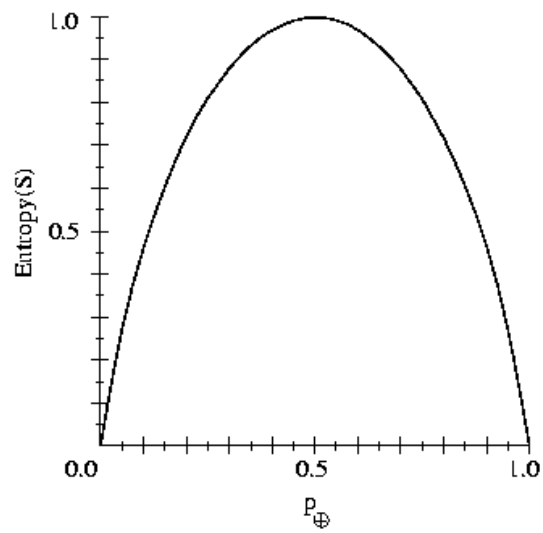
- Let S be a set of training examples.
 p_+ = proportion of positive examples.
 p_- = proportion of negative examples

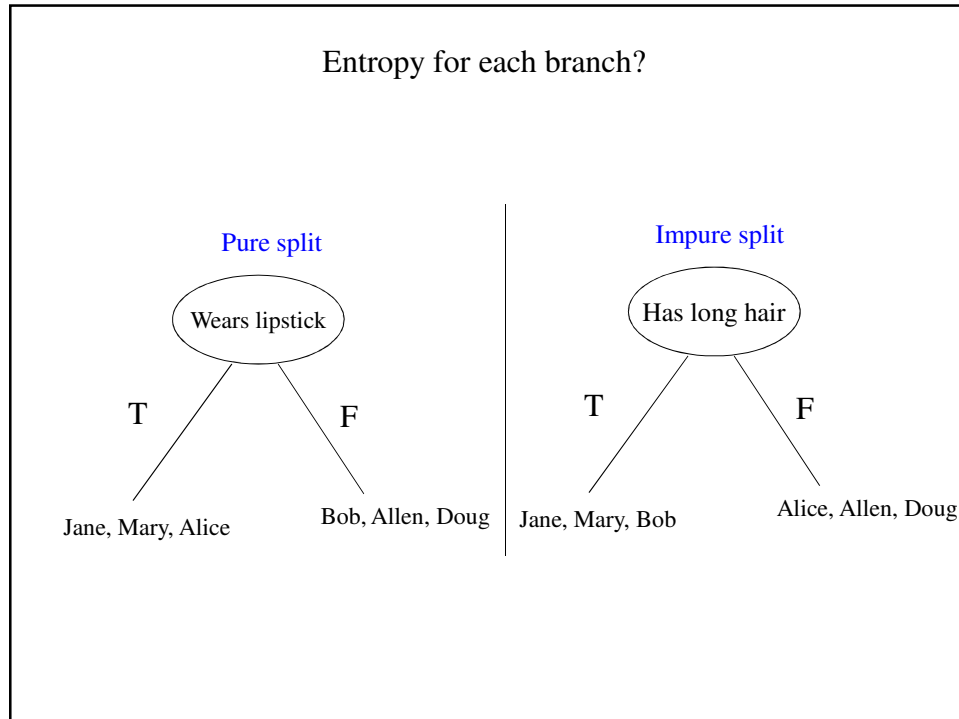
$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

- Entropy measures the degree of uniformity or non-uniformity in a collection.
- Roughly measures how predictable collection is, only on basis of distribution of + and - examples.

Entropy

- When is entropy zero?
- When is entropy maximum, and what is its value?





- What is the entropy of the given training set?

<u>Day</u>	<u>Outlook</u>	<u>Temp</u>	<u>Humidity</u>	<u>Wind</u>	<u>PlayTennis</u>
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

- Suppose you're now given a new example. In absence of any additional information, what classification should you guess?

- Entropy gives minimum number of bits of information needed to encode the classification of an arbitrary member of S .
 - If $p_+ = 1$, don't need any bits (entropy 0)
 - If $p_+ = .5$, need one bit (+ or -)
 - If $p_+ = .8$, can encode collection of $\{+,-\}$ values using on average less than 1 bit per value
 - Can you explain how we might do this?

- What is definition of entropy if each attribute can take on c different values?

Information gain

– For example, what if we know value of Humidity?

–

- If “high”, entropy is

$$\text{Entropy}(S_{\text{Hum=High}}) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.985$$

- If “normal”, entropy is

$$\text{Entropy}(S_{\text{Hum=Normal}}) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.592$$

- We can now calculate the average:

$$\begin{aligned} \text{Entropy}(S_{\text{Hum}}) &= \frac{|S_{\text{Hum=High}}|}{|S|} \text{Entropy}(S_{\text{Hum=High}}) + \frac{|S_{\text{Hum=Normal}}|}{|S|} \text{Entropy}(S_{\text{Hum=Normal}}) \\ &= \frac{1}{2} 0.985 + \frac{1}{2} 0.592 = 0.789 \end{aligned}$$

- This is lower than the entropy of S (0.940). What does this mean?

Information gain

- Intuitively entropy has gone down, so information has been gained by this partitioning. How much?
- Info gain = $E(S) - E(S_H) = 0.151$

Information gain

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

where $\text{Values}(A)$ is set of possible values for A , and

$$S_v = \{s \in S : A(s) = v\}$$

Information gain

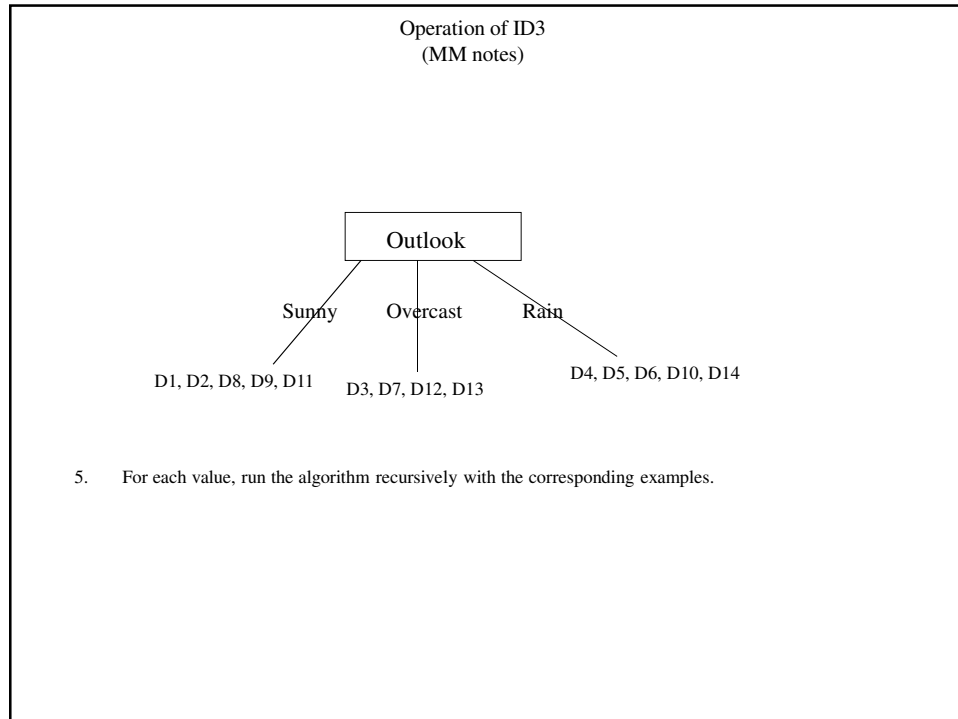
$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where $\text{Values}(A)$ is set of possible values for A , and

$$S_v = \{s \in S : A(s) = v\}$$

Operation of ID3

1. Compute information gain for each attribute.
 - Entropy of $S = .940$
 - Entropy of $S_{\text{Outlook}} = 5/14 E(\text{Out} = \text{Sunny}) + 4/14 E(\text{Out} = \text{Overcast}) + 5/14 E(\text{Out} = \text{Rain}) =$
 $5/14 (0.971) + 4/14 (0) + 5/14 (0.971) = 0.694$
Gain of $S_{\text{Outlook}} = 0.940 - 0.694 = 0.246$
 - Entropy of $S_{\text{Temperature}} = 4/14 E(\text{Temp}=\text{Hot}) + 6/14 E(\text{Temp}=\text{Mild}) + 4/14 E(\text{Temp}=\text{Cool}) =$
 0.911
Gain of $S_{\text{Temperature}} = 0.029$
 - Entropy of $S_{\text{Humidity}} = 1/2 E(\text{Hum} = \text{High}) + 1/2 E(\text{Hum} = \text{Normal}) = 0.789$
Gain of $S_{\text{Humidity}} = 0.151$
 - Entropy of $S_{\text{Wind}} = 0.892$
Gain of $S_{\text{Wind}} = 0.048$
2. Choose the best attribute A to be the root of the tree.
3. Add a branch for each value of A .
4. Sort training examples to each branch according to their value of A .



ID3's Inductive Bias

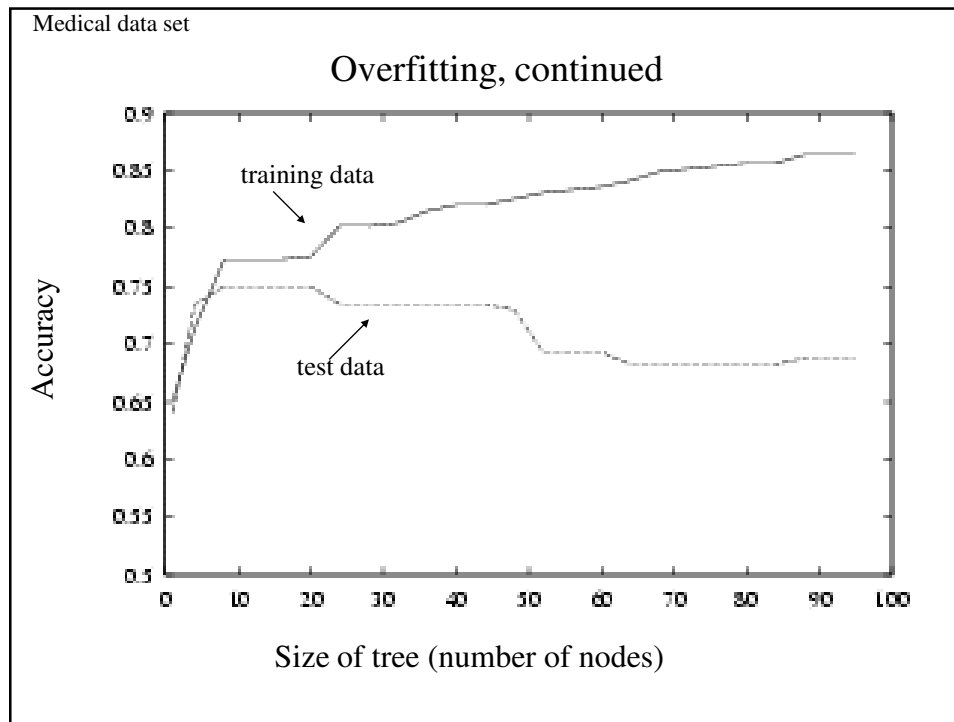
- Given a set of training examples, there are typically many decision trees consistent with that set.
 - E.g., what would be another decision tree consistent with the example training data?
- Of all these, which one does ID3 construct?
 - First acceptable tree found in greedy search

ID3's Inductive Bias, continued

- Algorithm does two things:
 - Favors shorter trees over longer ones
 - Places attributes with highest information gain closest to root.
- What would be an algorithm that explicitly constructs the shortest possible tree consistent with the training data?
- ID3: Efficient approximation to “find shortest tree” method

Overfitting

- Formal definition of **overfitting**:
 - Given a hypothesis space H , a hypothesis $h \in H$ is said to **overfit** the training data if there exists some alternative $h' \in H$, such that
$$\text{TrainingError}(h) < \text{TrainingError}(h'),$$
but
$$\text{TestError}(h') < \text{TestError}(h).$$



Overfitting, continued

- How to avoid overfitting:
 - Stop growing the tree early, before it reaches point of perfect classification of training data.
 - Allow tree to overfit the data, but then prune the tree.

Interesting observation about noise

- Quinlan found that adding a low level of noise to a training set resulted in trees with higher classification errors on training set, but with lower classification errors on new objects (not in training set).
- Explanation?
- Moral: “It is counter-productive to eliminate noise from the attribute information in the training set if these same attributes will be subject to high noise levels when the induced decision tree is put to use.”

Pruning the Tree

- Pruning:
 - Remove subtree below a decision node.
 - Create a leaf node there, and assign most common classification of the training examples affiliated with that node.
 - Helps get rid of nodes due to overfitting.

- Reduced-error pruning:
 - Consider each decision node as candidate for pruning.
 - For each node, try pruning node. Measure accuracy of pruned tree over validation set.
 - Select single-node pruning that yields best increase in accuracy over validation set.
 - If no increase, select one of the single-node prunings that does not decrease accuracy.
 - If all prunings decrease accuracy, then don't prune. Otherwise, continue this process until further pruning is harmful.

Continuous valued attributes

- Original decision trees: Two discrete aspects:
 - Target class (e.g., “*PlayTennis*”) has discrete values
 - Attributes (e.g., “*Humidity*”) have discrete values
- How to incorporate continuous-valued decision attributes?
 - E.g., *Humidity* \in [0,100]

Continuous valued attributes, continued

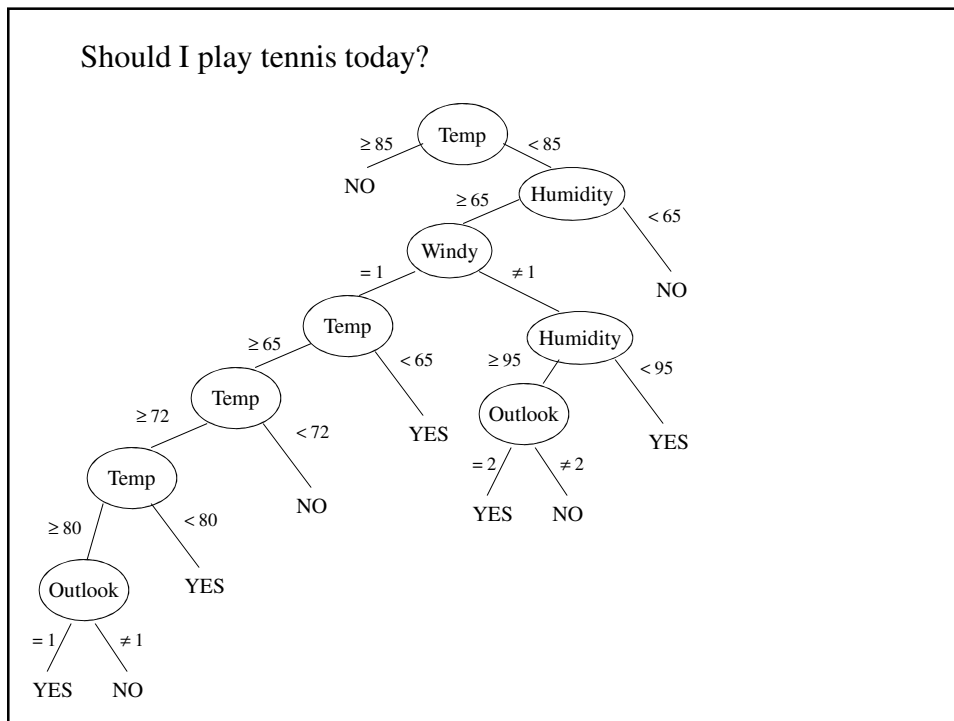
- Create new attributes, e.g., $Humidity_c$ true if $Humidity < c$, false otherwise.
- How to choose c ?
 - Find c that maximizes information gain.

Continuous valued attributes, continued

- Sort examples according to continuous value of $Humidity$

<i>Humidity:</i>	10	25	55	60	72	85
<i>PlayTennis:</i>	Yes	No	Yes	Yes	No	No
- Find adjacent examples that differ in target classification.
- Choose candidate c as midpoint of the corresponding interval.
 - Can show that optimal c must always lie at such a boundary.
- Then calculate information gain for each candidate c .
- Choose best one.
- Put new attribute $Humidity_c$ in pool of attributes.

Outlook	Temp	Humidity	Windy	Class	Training Set
1	85	85	0	0	
1	80	90	1	0	
2	83	78	0	1	
3	70	96	0	1	
3	68	80	0	1	
3	65	70	1	0	
2	64	65	1	1	
1	72	95	0	0	
1	69	70	0	1	
3	75	80	0	1	
1	75	70	1	1	
2	72	90	1	1	
2	81	75	0	1	
3	71	80	1	0	
2	82	70	1	0	
3	82	90	1	1	
1	76	92	0	1	
2	73	60	0	0	



Example: Spambase Database Features

word_freq_make:	continuous.	word_freq_your:	continuous.
word_freq_address:	continuous.	word_freq_font:	continuous.
word_freq_all:	continuous.	word_freq_000:	continuous.
word_freq_3d:	continuous.	word_freq_money:	continuous.
word_freq_our:	continuous.	word_freq_hp:	continuous.
word_freq_over:	continuous.	word_freq_hpl:	continuous.
word_freq_remove:	continuous.	word_freq_george:	continuous.
word_freq_internet:	continuous.	word_freq_650:	continuous.
word_freq_order:	continuous.	word_freq_lab:	continuous.
word_freq_mail:	continuous.	word_freq_labs:	continuous.
word_freq_receive:	continuous.	word_freq_telnet:	continuous.
word_freq_will:	continuous.	word_freq_857:	continuous.
word_freq_people:	continuous.	word_freq_data:	continuous.
word_freq_report:	continuous.	word_freq_415:	continuous.
word_freq_addresses:	continuous.	word_freq_85:	continuous.
word_freq_free:	continuous.	word_freq_technology:	continuous.
word_freq_business:	continuous.	word_freq_1999:	continuous.
word_freq_email:	continuous.	word_freq_parts:	continuous.
word_freq_you:	continuous.	word_freq_pm:	continuous.
word_freq_credit:	continuous.	word_freq_direct:	continuous.
		word_freq_cs:	continuous.
		word_freq_meeting:	continuous.

word_freq_original:	continuous.
word_freq_project:	continuous.
word_freq_re:	continuous.
word_freq_edu:	continuous.
word_freq_table:	continuous.
word_freq_conference:	continuous.
char_freq_;	continuous.
char_freq(:	continuous.
char_freq[_]	continuous.
char_freq_!	continuous.
char_freq_\$	continuous.
char_freq_#	continuous.
capital_run_length_average:	continuous.
capital_run_length_longest:	continuous.
capital_run_length_total:	continuous.

Spambase Database example entries

0,0,1.19,0,0,0,0,0,0,0,0,0,0,0,0,0,0.59,3.57,0.1.19,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.59,0,0,0,0,0,0,1,1,24,0
0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,5,0,2,0.11,0,0,3.178,62,89,1

C4.5 Demo

Alternative measures for selecting attributes

- Recall intuition behind information gain measure:
 - We want to choose attribute that does the most work in classifying the training examples by itself.
 - So measure how much information is gained (or how much entropy is decreased) if that attribute is known.

- However, information gain measure favors attributes with many values.
- Extreme example: Suppose that we add attribute “*Date*” to each training example. Each training example has a different date.
- What is its information gain?

Day	Outlook	Temp	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

- Date will be chosen as root of the tree.
- But of course the resulting tree will not generalize.

- Possible solution: Use only binary decision trees.
 - Disadvantage: Often hard to understand such trees.
 - Disadvantage: Can lead to large increase in computation, since gain of many possible thresholds has to be computed.

- Another possible solution: If your data has attribute with many values, use an alternate measure to select attributes for decision tree.
- Quinlan proposes “gain ratio”.
 - Define a term, called IV, that measures how uniformly instances are spread out over different values of an attribute:

$$IV(A) = -\sum_{i=1}^v \frac{p_i + n_i}{p + n} \log_2 \frac{p_i + n_i}{p + n}$$

$$= -\sum_{i=1}^v \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

= entropy of S with respect
to the values of attribute A

- This is **low** if instances are concentrated in a small number of values; **high** if instances are evenly spread out over values. Thus it serves as a penalty term for the latter case.