UNSW
AUSTRALIA

Tree Learning

Never Stand Still

COMP9417 Machine Learning & Data Mining

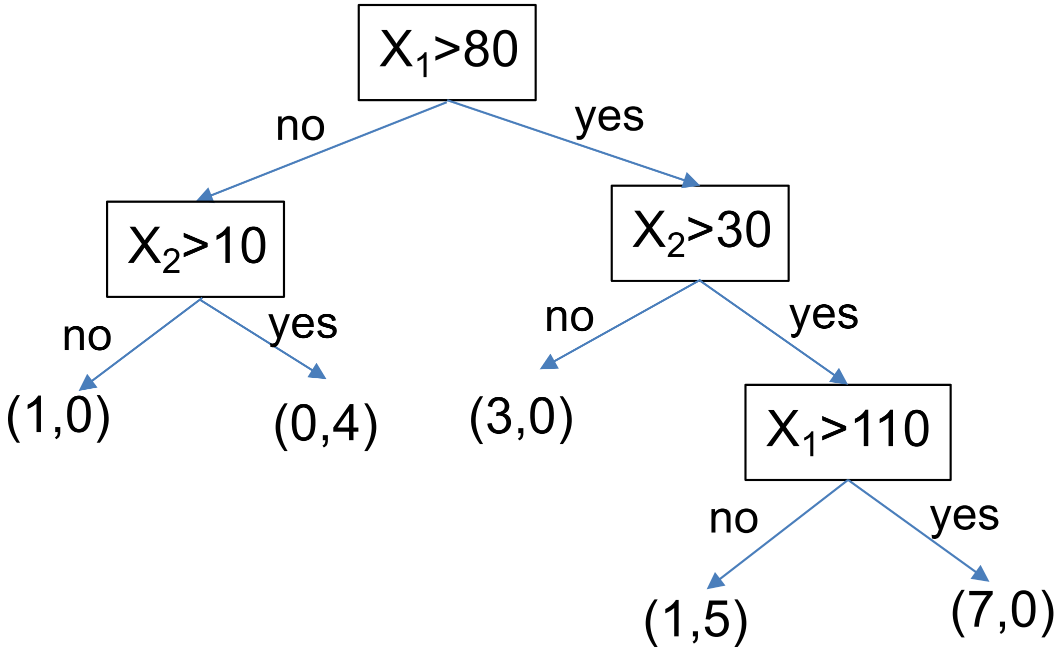
Aims

This lecture will enable you to describe decision tree learning, the use of entropy and the problem of overfitting. Following it you should be able to:

- Define the decision tree representation
- Reproduce the basic top-down algorithm for decision tree induction (TDIDT)
- Define entropy in the context of learning a classifier from examples
- Define overfitting in decision trees and how to control it
- Describe the inductive bias of the basic TDIDT algorithm
- Describe regression trees

Classification: Decision trees

Example: Let's go back to the fish example with two types of "*salmon*" and "*sea bass*" and assume we have two features $length$ and $width$ to classify fish type



$X_1>80$

no

yes

$X_2>10$

no

yes

(1,0)

(0,4)

$X_2>30$

no

yes

(3,0)

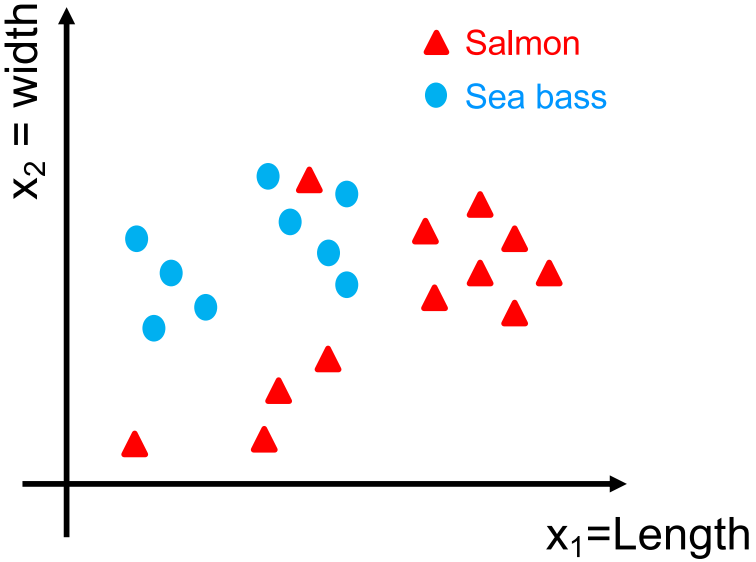
$X_1>110$

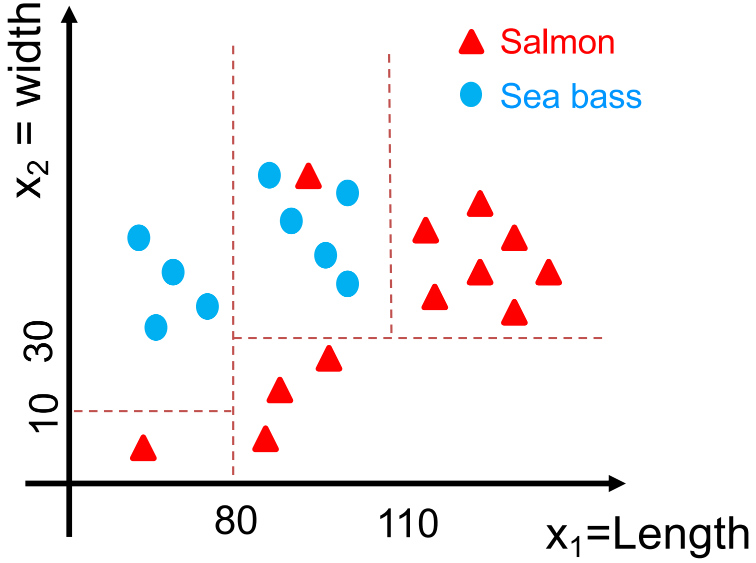
no

yes

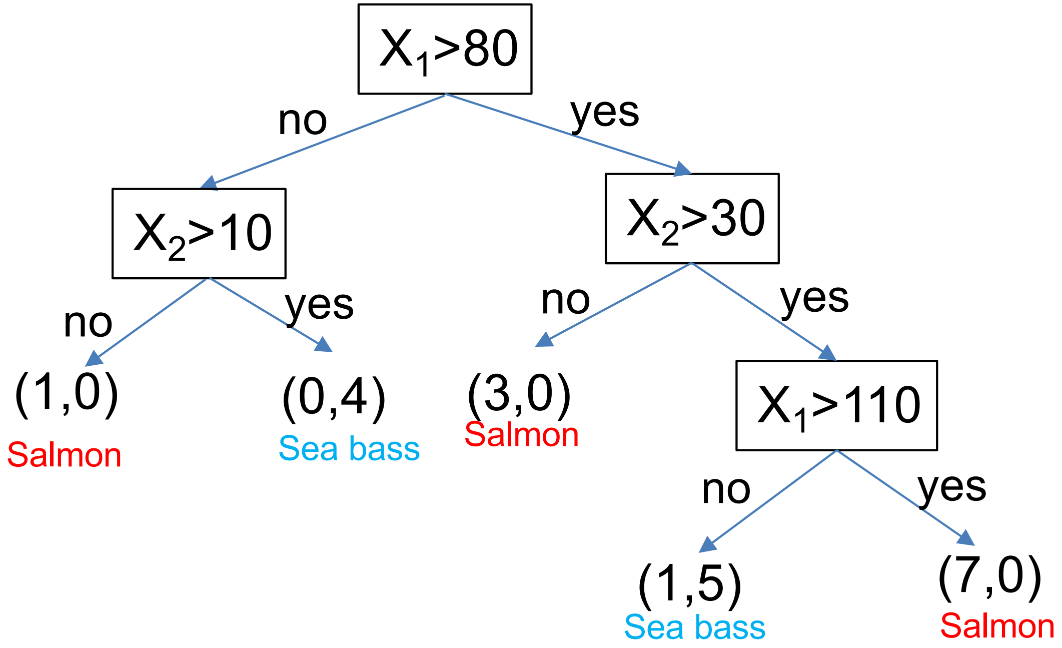
(1,5)

(7,0)





Classification: Decision trees



For any new sample, we start from the top of the tree and answer the questions and follow the appropriate road to the bottom and then decide about the label

Why use decision trees?

- Decision trees are probably the single most popular data mining tool
  - Easy to understand
  - Easy to implement
  - Easy to use
  - Computationally cheap (efficient, even on big data)
- There are some drawbacks, though — e.g., high variance
- They do classification, i.e., predict a categorical output from categorical and/or real inputs, or regression

Decision Tree for PlayTennis

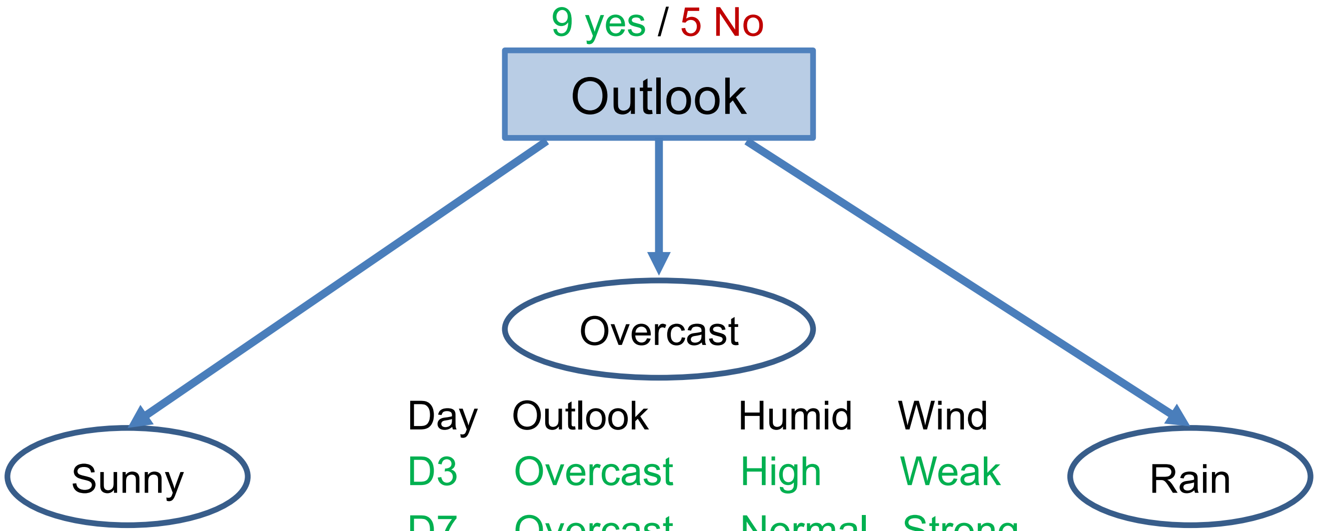
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Decision Tree for PlayTennis

Decision tree works in a divide and conquer fashion:

- Split into subsets
- Are they pure?
  - If yes: stop
  - If not: repeat
- For a new data, find the subset the data falls into

Decision Tree for PlayTennis



9 yes / 5 No

Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

2 yes / 3 No

Split Further

Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

4 yes / 0 No

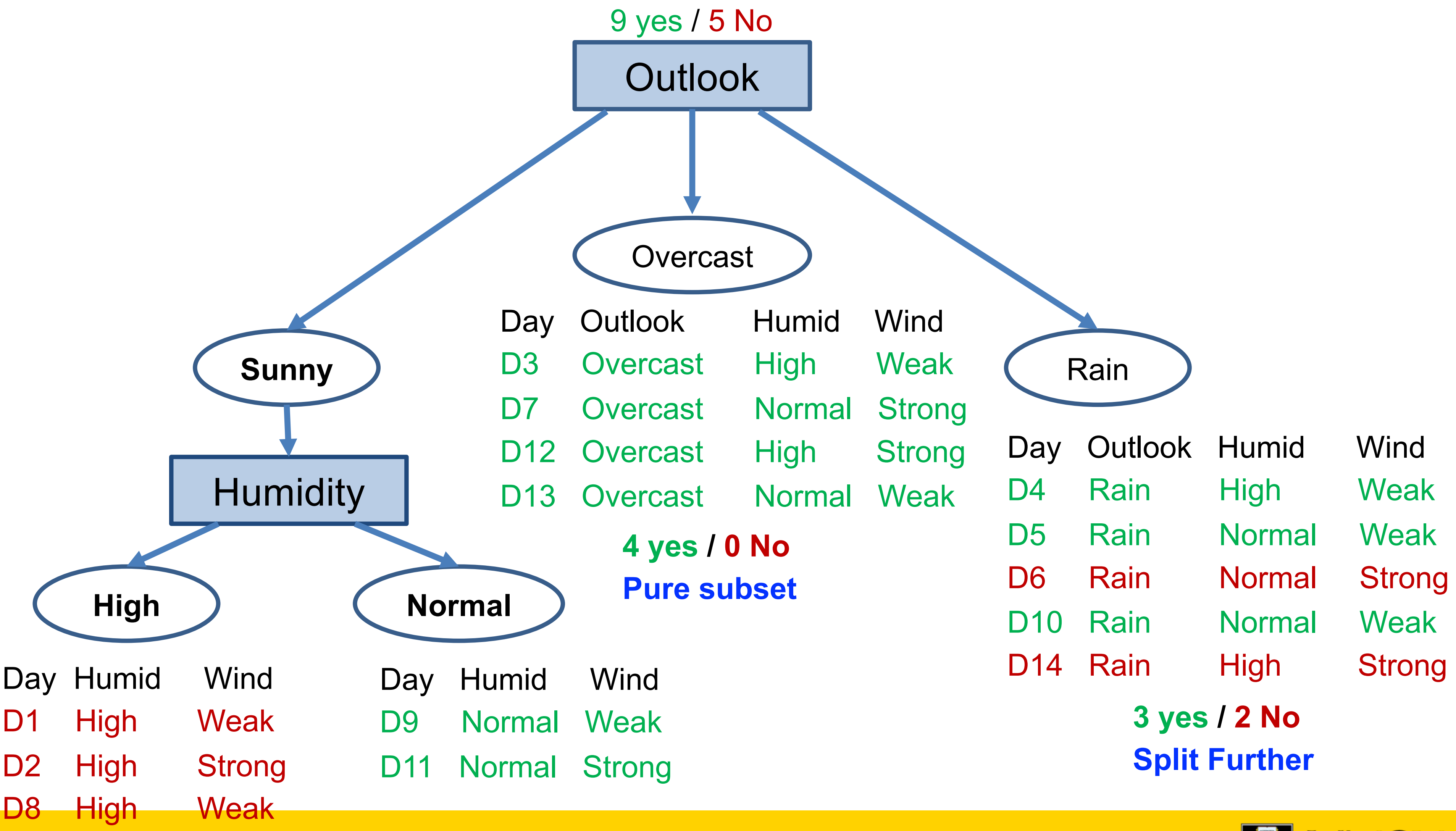
Pure subset

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3 yes / 2 No

Split Further

Decision Tree for PlayTennis



9 yes / 5 No

Outlook

Overcast

Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

4 yes / 0 No

Pure subset

Sunny

Humidity

High

Day	Humid	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Normal

Day	Humid	Wind
D9	Normal	Weak
D11	Normal	Strong

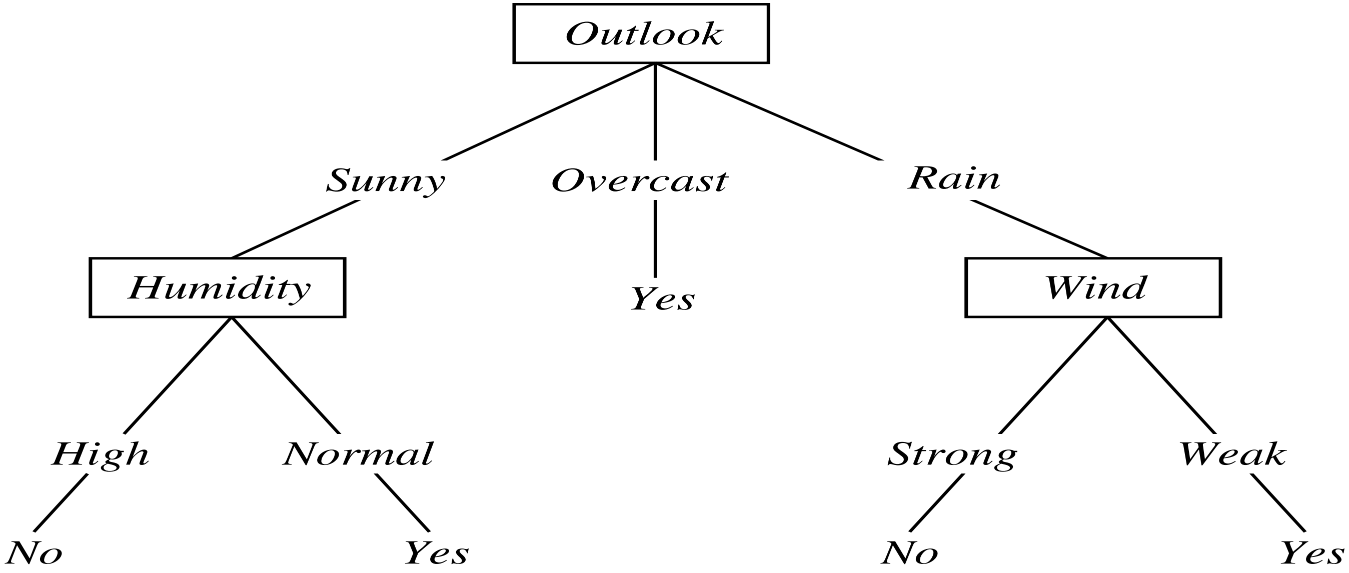
Rain

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3 yes / 2 No

Split Further

Decision Tree for PlayTennis



What would be the decision for a new data:

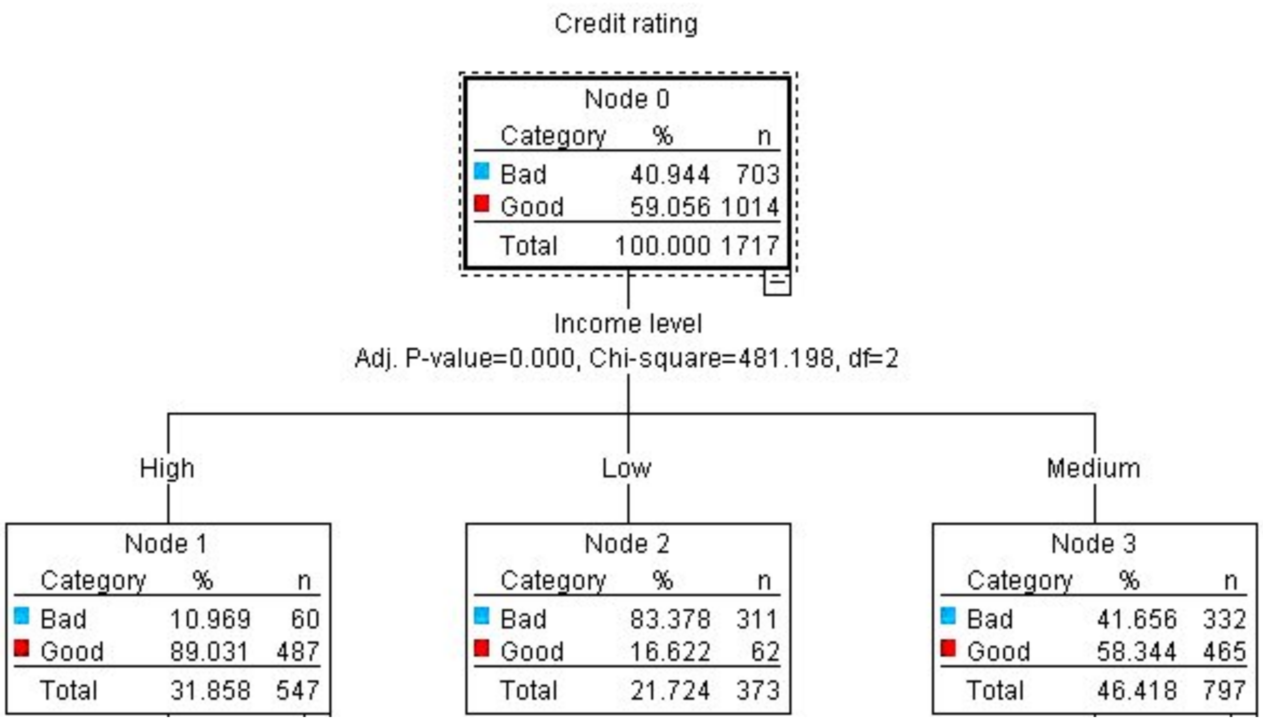
<Outlook = Rain, Humidity = High , Wind = Weak> ?

A Tree to Predict C-Section Risk

Learned from medical records of 1000 women. Negative examples are C-sections

```
[833+,167-] .83+ .17-
Fetal_Presentation = 1: [822+,116-] .88+ .12-
| Previous_Csection = 0: [767+,81-] .90+ .10-
| | Primiparous = 0: [399+,13-] .97+ .03-
| | Primiparous = 1: [368+,68-] .84+ .16-
| | | Fetal_Distress = 0: [334+,47-] .88+ .12-
| | | | Birth_Weight < 3349: [201+,10.6-] .95+ .05-
| | | | Birth_Weight >= 3349: [133+,36.4-] .78+ .22-
| | | Fetal_Distress = 1: [34+,21-] .62+ .38-
| Previous_Csection = 1: [55+,35-] .61+ .39-
Fetal_Presentation = 2: [3+,29-] .11+ .89-
Fetal_Presentation = 3: [8+,22-] .27+ .73-
```

Decision Tree for Credit Rating



Credit rating

Node 0

Category	%	n
🟦 Bad	40.944	703
🟥 Good	59.056	1014
Total	100.000	1717

Income level
Adj. P-value=0.000, Chi-square=481.198, df=2

High

Node 1

Category	%	n
🟦 Bad	10.969	60
🟥 Good	89.031	487
Total	31.858	547

Low

Node 2

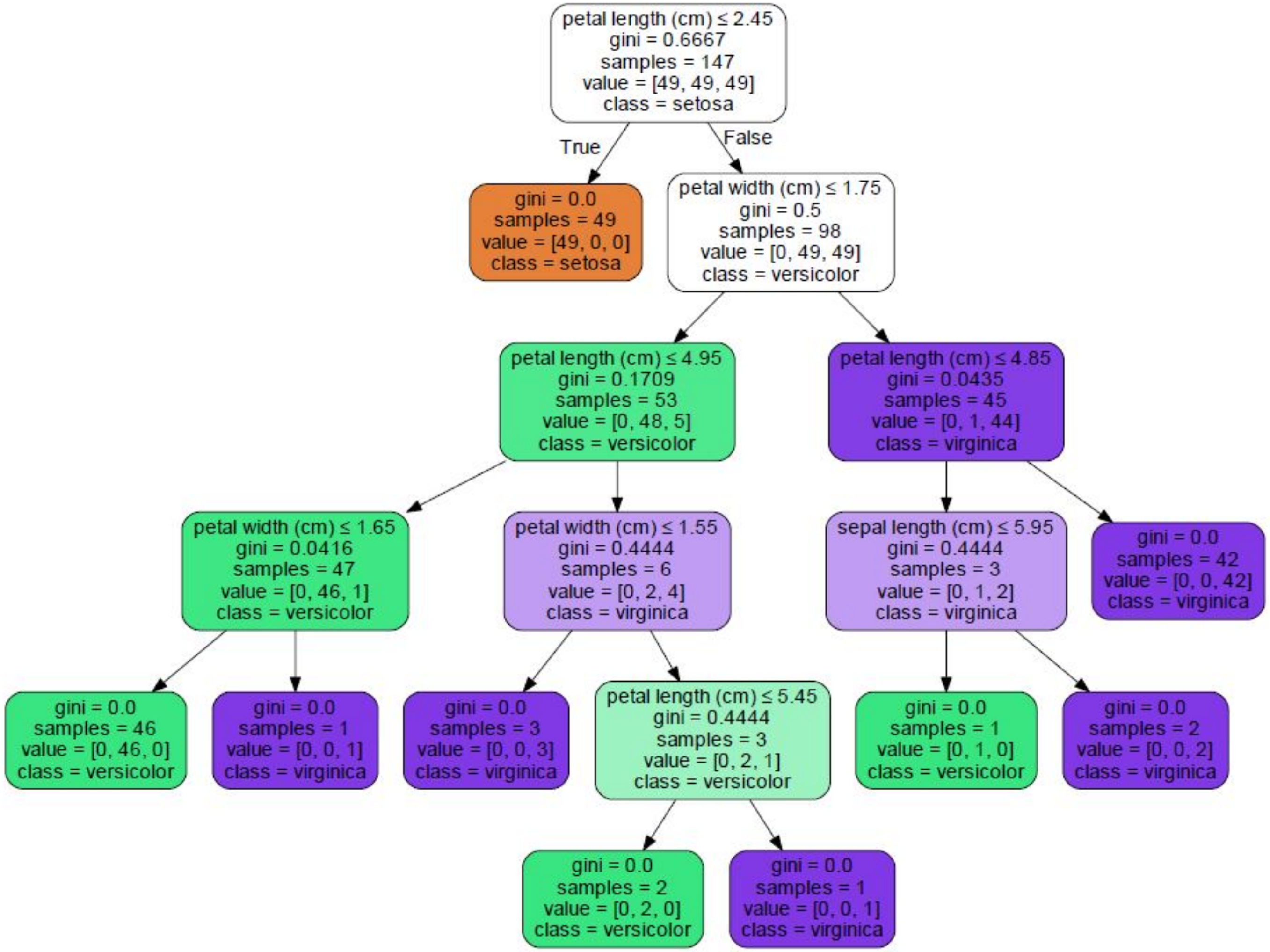
Category	%	n
🟦 Bad	83.378	311
🟥 Good	16.622	62
Total	21.724	373

Medium

Node 3

Category	%	n
🟦 Bad	41.656	332
🟥 Good	58.344	465
Total	46.418	797

Decision Tree for Fisher’s Iris data



Decision Trees

Decision tree representation:

- Each internal node tests an attribute
- Each branch corresponds to attribute value
- Each leaf node assigns a classification

Decision Tree: Expressiveness

Using decision trees to represent Boolean functions like AND (∧), OR (∨), XOR (⊕)

$\boldsymbol{X \wedge Y}$

$X = t$:

| $Y = t$: $true$

| $Y = f$: $false$

$X = f$: $false$

$\boldsymbol{X \vee Y}$

$X = t$: $true$

$X = f$:

| $Y = t$: $true$

| $Y = f$: $false$

Decision Tree: Expressiveness

$\boldsymbol{X \oplus Y}$

$X = t$:

$| \ Y = t: false$

$| \ Y = f: true$

$X = f$:

$| \ Y = t: true$

$| \ Y = f: false$

"In general, 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, and the tree itself to a disjunction of these conjunctions" (Mitchell, 1997)

When to Consider Decision Trees?

- Instances described by a mix of numeric features and discrete attribute–value pairs
- Target function is discrete valued (otherwise use regression trees)
- Possibly noisy training data
- Interpretability is an advantage

Examples are extremely numerous, including:

- Equipment or medical diagnosis
- Credit risk analysis
- Modelling calendar scheduling preferences
- etc.

Top-Down Induction of Decision Trees (TDIDT)

Main loop:

- $A \leftarrow$ the “best” decision attribute for next node to split examples
- Assign A as decision attribute for node
- For each value of A, create new descendant of node (child node)
- Split training examples to child nodes
- If training examples perfectly classified (pure subset), Then $STOP$, Else iterate over new child nodes

Discovered by two people independently:

- Ross Quinlan (ID3: 1986), (C4.5: 1993)
- Breiman et. al (CaRT: 1984) from statistics

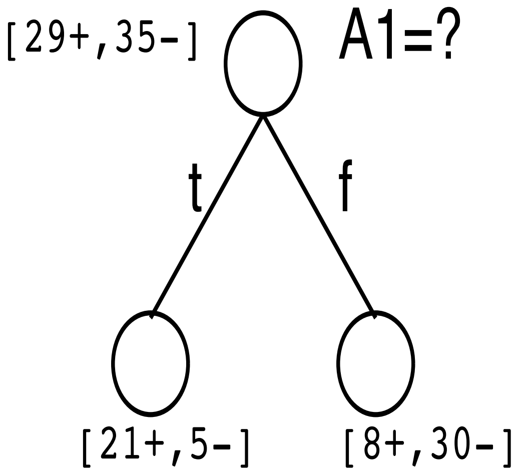
Top-Down Induction of Decision Trees (TDIDT)

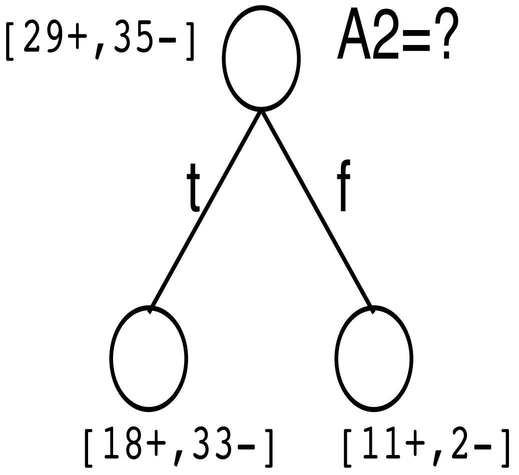
- ID3 (Iterative Dichotomiser 3)
  - Uses categorical attributes/features
  - For categorical targets
  - Is precursor of C4.5 (without restriction of categorical attributes)

- CaRT (Classification & Regression Tree)
  - Uses both categorical and numerical attributes/features
  - Supports both categorical & numerical targets

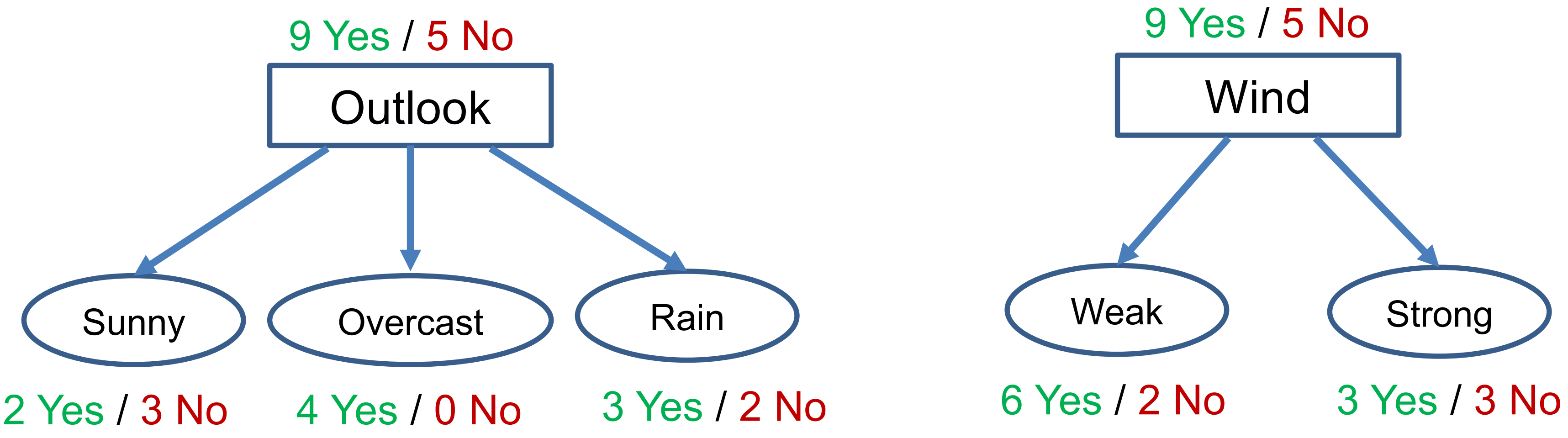
Here, the focus is on ID3 for classification.

Which attribute is best?

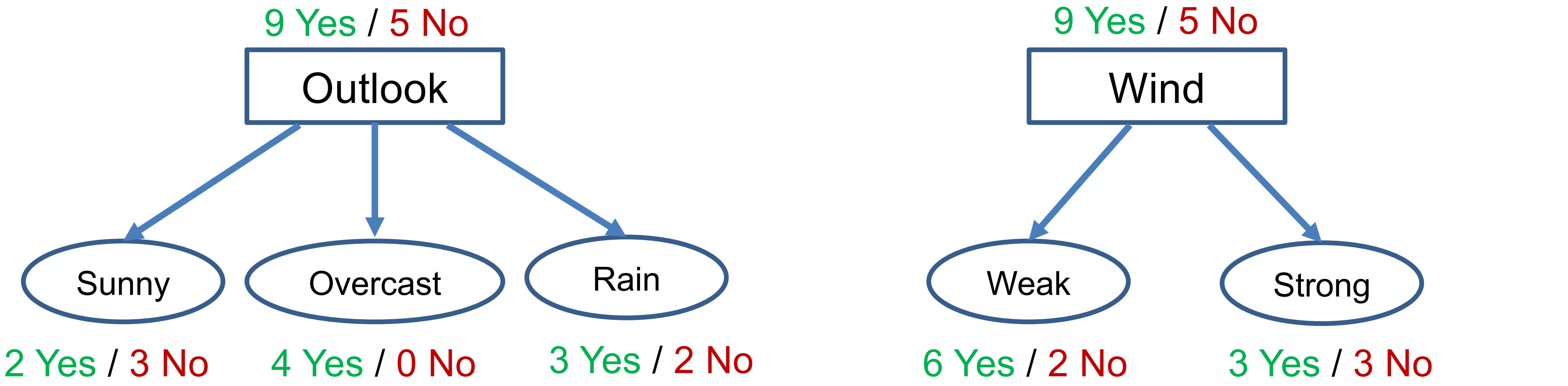




Which attribute is best?



Which attribute is best?



- We are looking for a split with higher “purity”
- We need to measure the purity of the split
    - More certain about our classes after split
        - A set with all examples belonging to one class is 100% pure
        - A set with 50% examples in one class and 50% in the other is 100% uncertain and impure

Which attribute is best?

There are many different ways to measure the purity of a subset, but one good measure for it is Entropy.

Entropy:

- From statistical point of view: is a measure of uncertainty
- From information theory point of view: The amount of information (in the Shannon sense) needed to specify the full state of a system

Entropy

Entropy measures the “impurity” of S

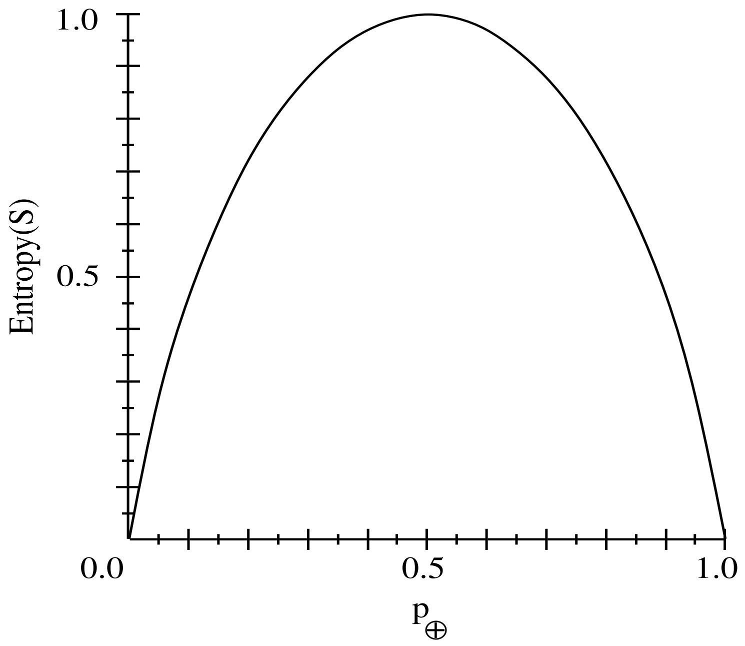
$$Entropy(S) = H(S) = -p_{\oplus}\ \log_2 p_{\oplus} - p_{\ominus}\ \log_2 p_{\ominus}$$

- S is subset of training examples
- $p_{\oplus}, p_{\ominus}$ are the portion(%) of positive and negative examples in S

Interpretation: if item x belongs to S, how many bits are needed to tell if x is positive or negative?

A “pure” sample set is one in which all examples are of the same class.

Entropy



Where:

S is a sample of training examples

$p_{\oplus}$is the proportion of positive examples in S

$p_{\ominus}$ is the proportion of negative examples in S

General Case

From information theory, the optimal number of bits to encode a symbol with probability p is $-\log_2 p$ ...

Suppose X can have one of k values ... $v_1, v_2, \dots, v_k$

$$P(X = v_1) = p_1, P(X = v_2) = p_2, \dots, P(X = v_k) = p_k$$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution ? It's

$$H(X) = -p_1 \ \log_2 p_1 - p_2 \ \log_2 p_2 - \cdots - p_k \ \log_2 p_k$$

$$= -\sum_{j=1}^{k} p_j \ \log_2 p_j$$

$H(X)$ is the *entropy* of X

Entropy

$Entropy(S)$ is the expected number of bits needed to encode class ($\oplus$ or $\ominus$) of randomly drawn member of S (under the optimal, shortest-length code)

So, expected number of bits to encode $\oplus$ or $\ominus$ of random member of S:

$$p_{\oplus}\ (-\log_2 p_{\oplus}) + p_{\ominus}\ (-\log_2 p_{\ominus})$$

$$Entropy(S) = -p_{\oplus}\ \log_2 p_{\oplus} - p_{\ominus}\ \log_2 p_{\ominus}$$

And we want to minimize this entropy.

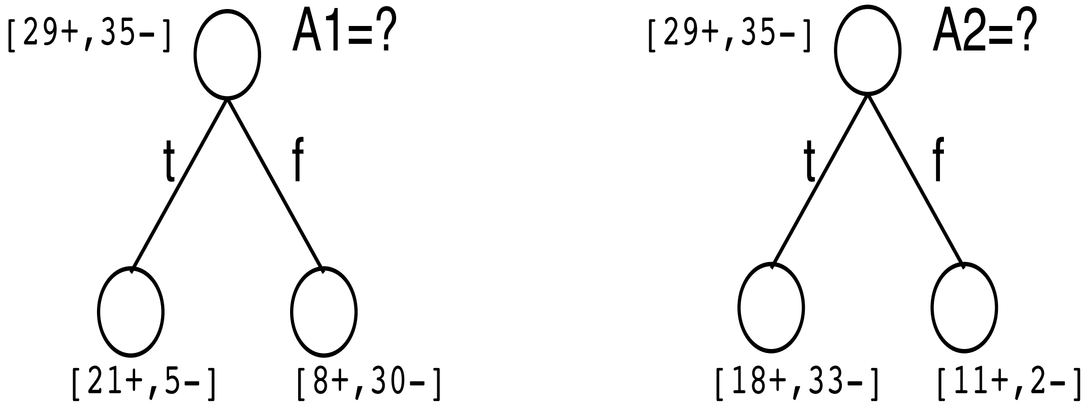
Entropy

- “High entropy”/”impure set” means X is very uniform and boring
    - Example: (3 samples from $\oplus$, 3 samples from $\ominus$)
    - $H(S) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$ (can be interpreted as 1 bits)

- “Low entropy”/”pure set” means X is not uniform and interesting
    - Example: (6 samples from $\oplus$, 0 samples from $\ominus$)
    - $H(S) = -\frac{6}{6}\log_2\frac{6}{6} - \frac{0}{6}\log_2\frac{0}{6} = 0$ (can be interpreted as 0 bits)

Entropy

Entropy is a measure in just one subset.



$$H(S) = -\frac{29}{64}\log_2\frac{29}{64} - \frac{35}{64}\log_2\frac{35}{64} = 0.9936$$

But, we are interested in expected drop in entropy after split, to decide which attribute is the best. How?

Information Gain

- $Gain(S, A)$ is the expected reduction in entropy due to sorting on A

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

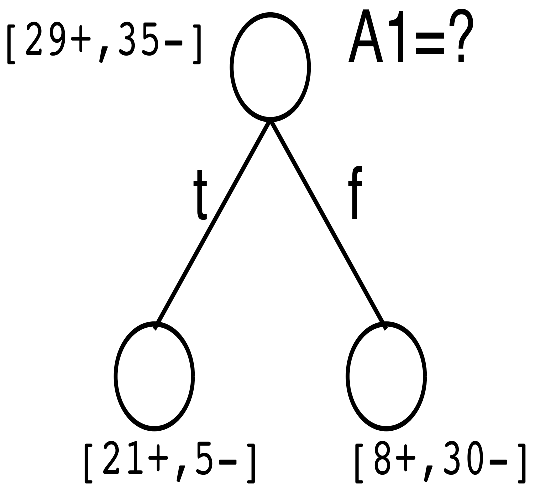
  - ○ v is the possible values of attribute A
  - ○ S is the set of examples we want to split
  - ○ S_v is the subset of examples where $X_A = v$

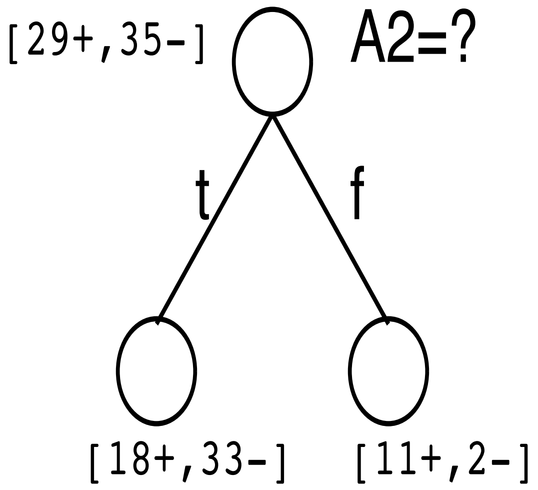
- We want to find the attribution which maximizes the gain

This is also called “mutual information” between attribute A and class labels of S

Information Gain

What is the information gain for attribute $A1$ and $A2$?





Information Gain

$$Gain(S, A1) = Entropy(S) - (\frac{|S_t|}{|S|} Entropy(S_t) + \frac{|S_f|}{|S|} Entropy\left(S_f\right))$$

$$= \ 0.9936 \ - ((\frac{26}{64}(-\frac{21}{26}\log_2\left(\frac{21}{26}\right) - \frac{5}{26}\log_2(\frac{5}{26}))) +$$

$$(\frac{38}{64}(-\frac{8}{38}\log_2\left(\frac{8}{38}\right) - \frac{30}{38}\log_2(\frac{30}{38}))))$$

$$= 0.9936 - (\ 0.2869 + 0.4408)$$

$$= 0.2658$$

$$Gain(S, A2) = 0.9936 - (\ 0.7464 + 0.0828)$$

$$= 0.1643$$

Information Gain

So in this example, we choose $A1$, since it gives a larger expected reduction in entropy.

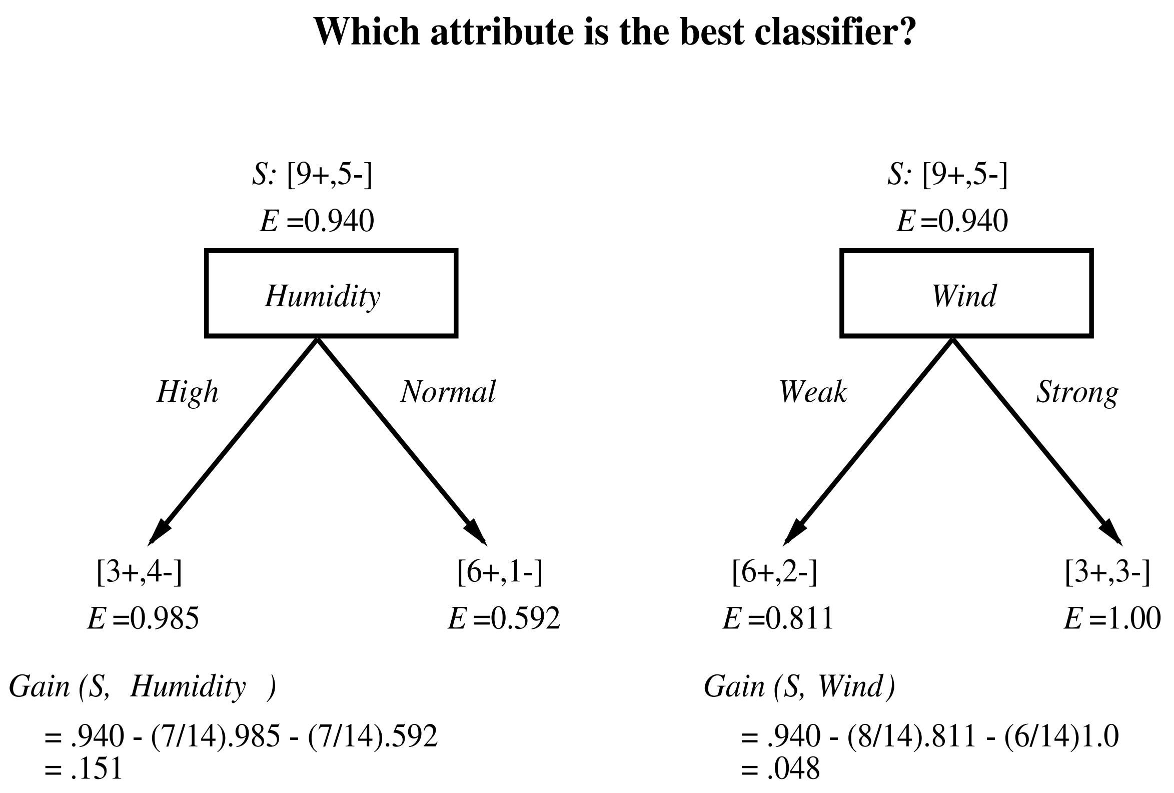
To select best attribute at each branch:

- take every/remaining attributes in your data
- Compute information gain for that attribute
- Select the attribute that has the highest information gain

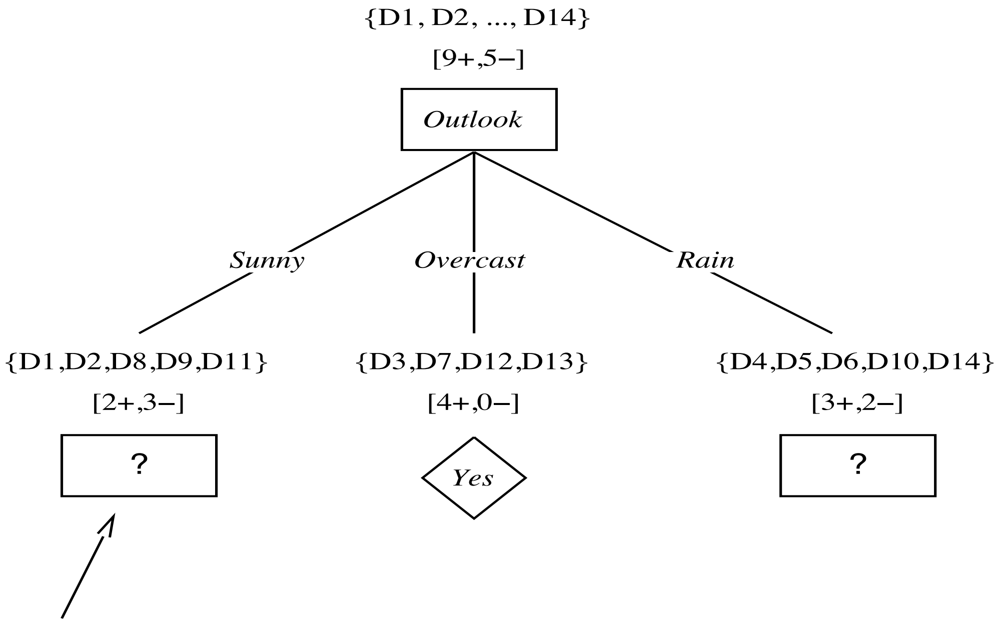
Training Examples

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

Information gain once more



Information gain once more



{D1, D2, ..., D14}

[9+,5−]

Outlook

Sunny

Overcast

Rain

{D1,D2,D8,D9,D11}

[2+,3−]

?

{D3,D7,D12,D13}

[4+,0−]

Yes

{D4,D5,D6,D10,D14}

[3+,2−]

?

Which attribute should be tested here?

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

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

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

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

Information Gain

Will information gain always find the prefect solution?

Information Gain Limitations

- Information gain is more biased towards attributes with large number of values/categories
    - Subsets are more likely to be pure if there is a large number of values
    - This can result overfitting (selection of an attribute which is non-optimal for prediction) which does not generalize well to unseen data
- Suggested solution: gain ratio
    - A modification of information gain that reduces the bias
    - Takes number and size of branches into account

Gain ratio

$$SplitEntropy(S,A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

Where:

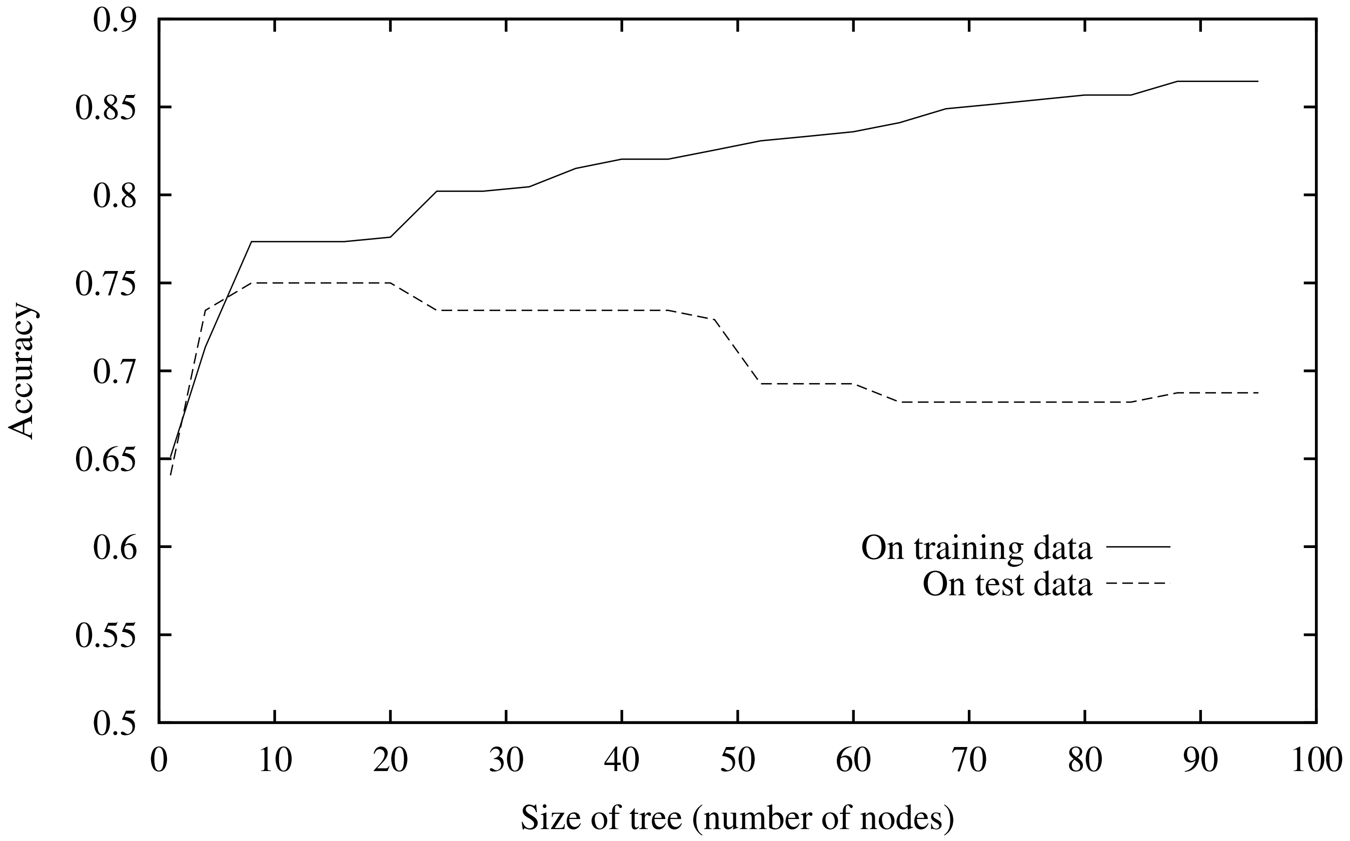
- A: candidate attribute
- v: possible values of A
- S: Set of examples $\{X\}$ at the node
- S_v: subset where $X_A = v$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitEntropy(S,A)}$$

Overfitting in Decision Trees

- Can always classify training examples perfectly
  - If necessary, keep splitting until each node contains 1 example
  - Singleton subset (leaf nodes with one example) which is by definition pure
- But this is not always ideal
  - In leaf nodes with singleton subsets, you have no confidence in your decision

Overfitting in Decision Tree Learning



Overfitting in General

Consider error of hypothesis h over

- training data: $error_{train}(h)$
- entire distribution $\mathcal{D}$ of data: $error_{\mathcal{D}}(h)$

Definition

Hypothesis $h \in H$ overfits training data if there is an alternative hypothesis $h' \in H$ such that

$$error_{train}(h) < error_{train}(h')$$

And

$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$

Avoiding Overfitting

How can we avoid overfitting? **Pruning**

- **pre-pruning:** stop growing when data split not statistically significant
- **post-pruning:** grow full tree, then remove sub-trees which are overfitting (based on validation set)

Post-pruning avoids problem of “early stopping”

Avoiding Overfitting

Pre-pruning

- Can be based on statistical significance test
- Stops growing the tree when there is no statistically significant association between any attribute and the class at a particular node
- For example, in ID3: chi-squared test plus information gain
  - only statistically significant attributes were allowed to be selected by information gain procedure

Avoiding Overfitting

Pre-pruning

- Simplest approach: stop growing the tree when fewer than some lower-bound on the number of examples at a leaf
  - In *sklearn*, this parameter is $\min\_samples\_leaf$
- Stop when Entropy changes is smaller than a lower-round
  - In *sklearn*, the parameter $\min\_impurity\_decrease$ enables stopping when this falls below a lower-bound

Avoiding Overfitting

- Other pre-pruning approaches:
  - Maximum number of leaf nodes
    - in *sklearn*, $\max\_leaf\_nodes$
  - Minimum number of samples to split
    - in *sklearn* $\min\_samples\_split$
  - Maximum depth
    - in *sklearn*, $\max\_depth$

Avoiding Overfitting

Early stopping

- Pre-pruning may suffer from early stopping: may stop the growth of tree prematurely
- Classic example: XOR/Parity-problem
  - No individual attribute exhibits a significant association with the class
  - Target structure only visible in fully expanded tree
  - Pre-pruning won't expand the root node
- But: XOR-type problems not common in practice
- And: pre-pruning faster than post-pruning

Avoiding Overfitting

Post-pruning

- Builds full tree first and prunes it afterwards
    - Attribute interactions are visible in fully-grown tree
- Problem: identification of subtrees and nodes that are due to chance effects
- Subtree replacement
- Possible strategies: error estimation, significance testing, MDL principle. We examine
    - reduced-error Pruning
    - Minimum error
    - Smallest tree

Reduced-Error Pruning

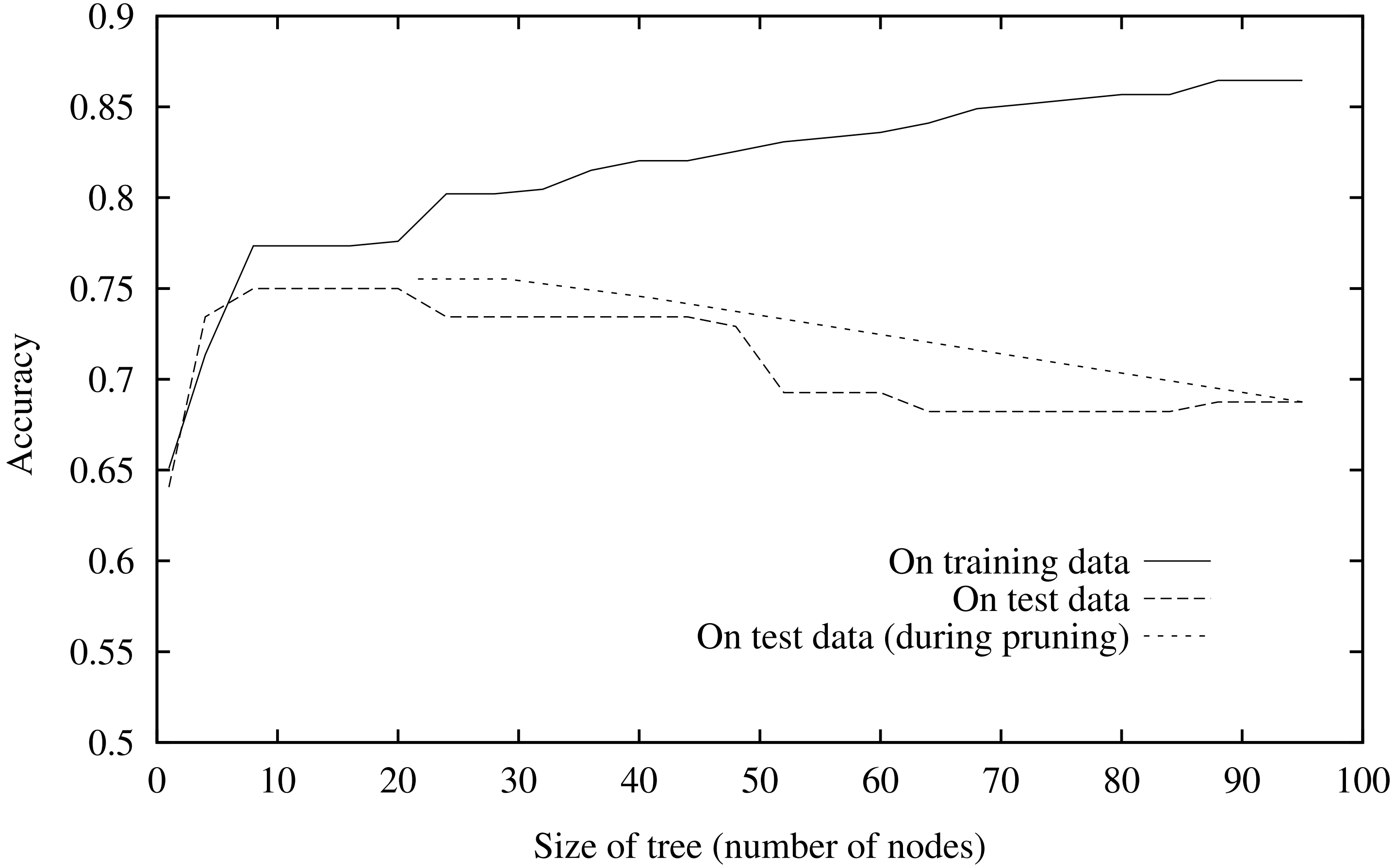
Split data into *training* and *validation* set

Do until further pruning is harmful:

- Evaluate the impact of pruning nodes in the bottom of the tree and replacing them with a leaf node on *validation* set
- Greedily remove the one that most improves *validation* set accuracy

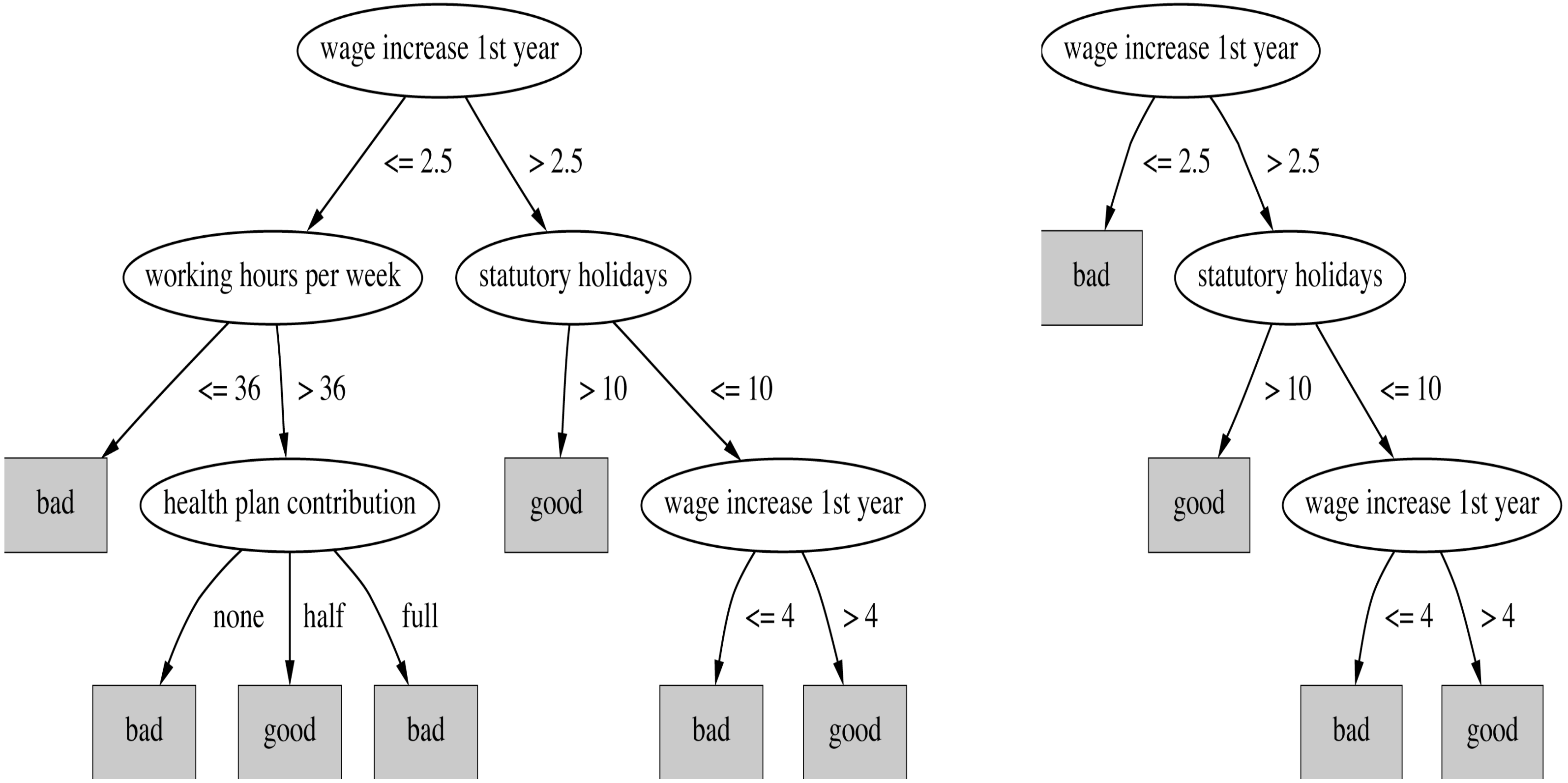
- **Good:** because it is simple and produces smallest version of most accurate subtree
- **Not so good:** because it reduces effective size of training set

Effect of Reduced-Error Pruning



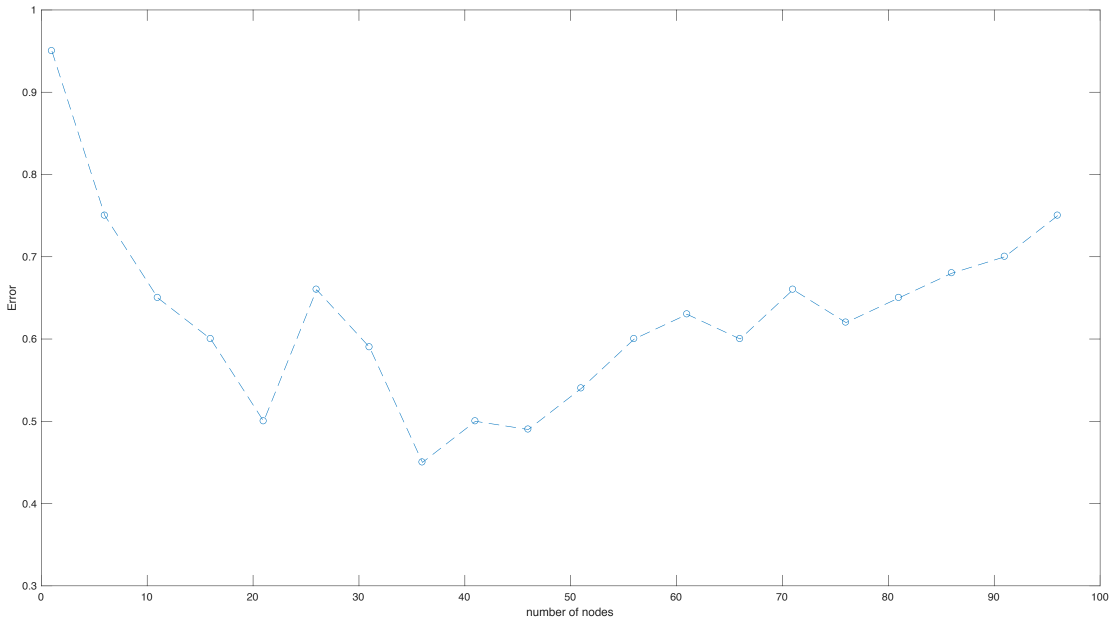
Pruning operator: Sub-tree replacement

Bottom-up:
Tree is considered for replacement once all its sub-trees have been considered



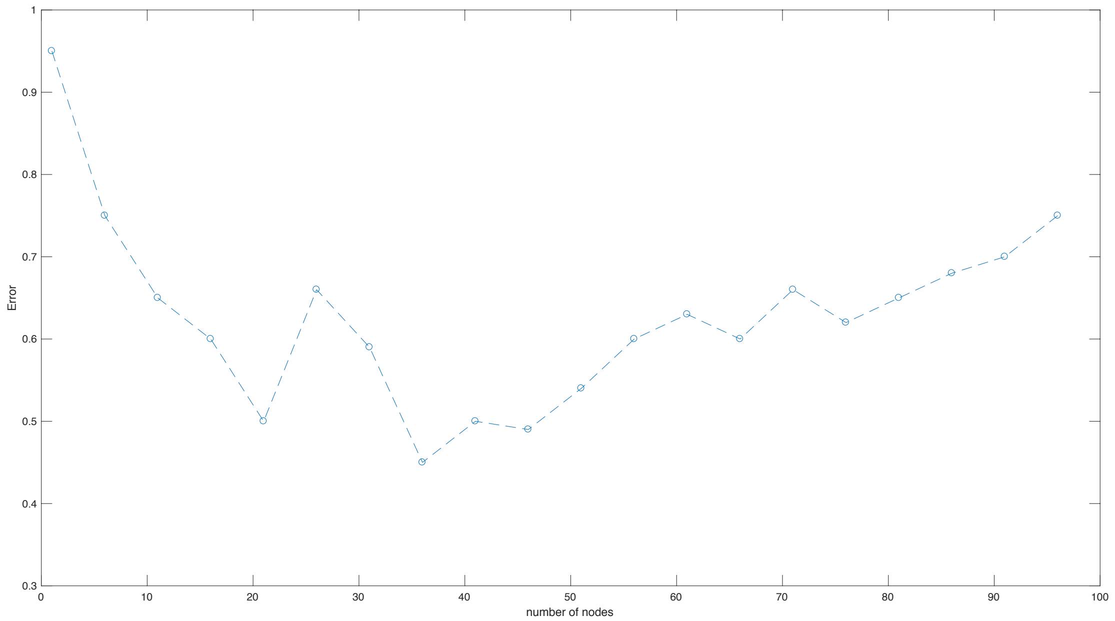
Minimum Error

- Keep the Cross-Validation error during the growing
- Prune back to the point where the cross-validated error is a minimum.
  - In the following example, the first 36 nodes will be kept

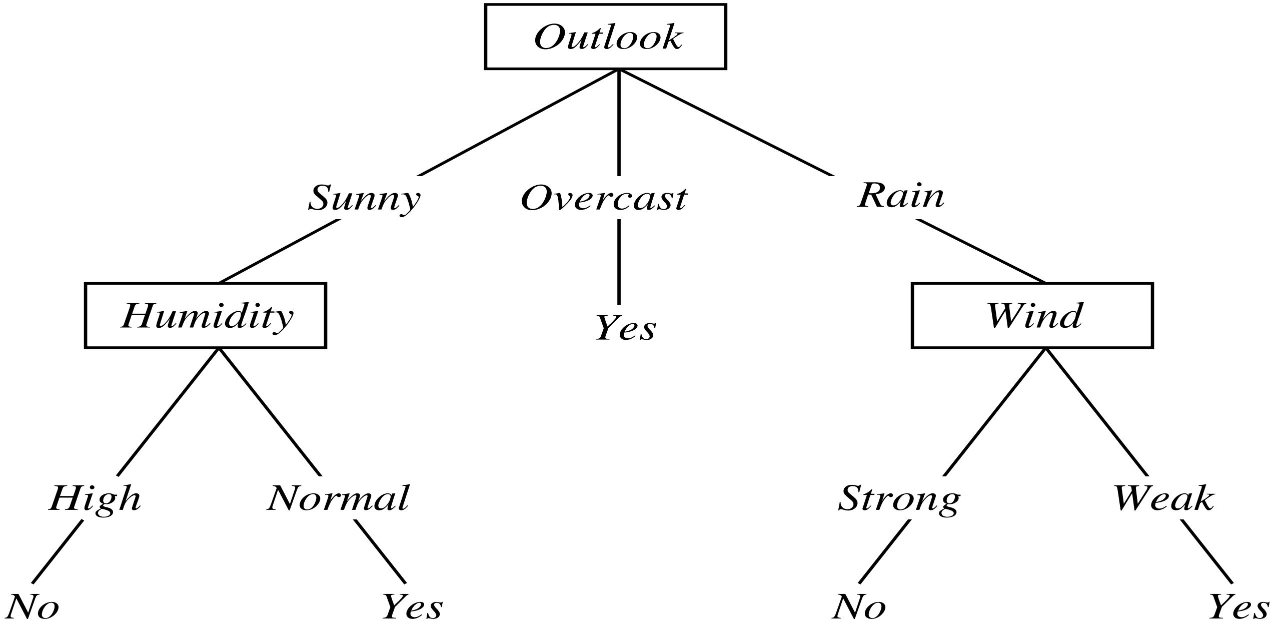


Smallest Tree

- Keep the Cross-Validation error during the growing
- Prune back the tree to the smallest tree within 1 standard error of the minimum error.
  - In the following example, the first 20 nodes will be kept



Converting A Tree to Rules



If $(Outlook = Sunny) \wedge (Humidity = High)$
Then $PlayTennis = No$

If $(Outlook = Sunny) \wedge (Humidity = Normal)$
Then $PlayTennis = Yes$
…

Rules from Trees

Rules can be simpler than trees but just as accurate, e.g.:

- path from root to leaf (in unpruned tree) forms a rule
    - i.e., tree forms a set of rules
- can simplify rules independently by deleting conditions (by pruning)
    - i.e., rules can be generalized while maintaining accuracy

Continuous Valued Attributes

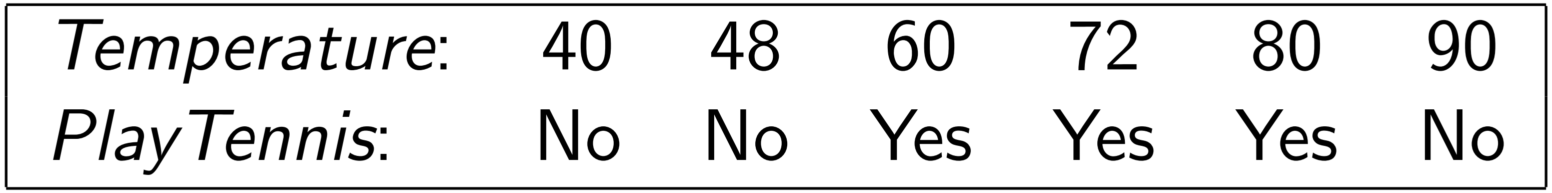
Decision trees originated for discrete attributes only. Now: continuous attributes.

Can create a discrete attribute to test continuous value:

- $Temperature = 82.5$
- $(Temperature > 72.3) \in \{t, f\}$
- Usual method: continuous attributes have a binary split
- Note:
  - discrete attributes – one split exhausts all values
  - continuous attributes – can have many splits in a tree

Continuous Valued Attributes

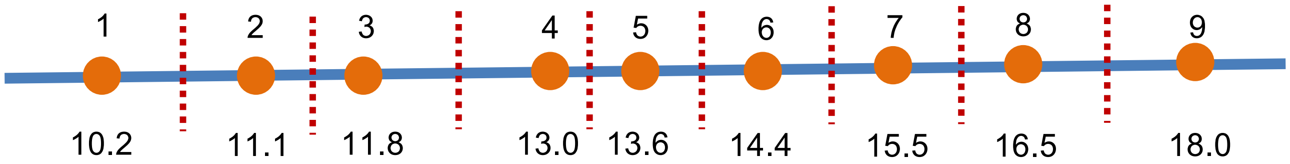
- Splits evaluated on all possible split points
- More computation: $n - 1$ possible splits for n values of an attribute in training set
- Fayyad (1991)
  - sort examples on continuous attribute
  - find midway boundaries where class changes, e.g. for $Temperature$ $\frac{(48+60)}{2}$ and $\frac{(80+90)}{2}$
- Choose best split point by info gain (or evaluation of choice)
- Note: C4.5 uses actual values in data



Temperature:	40	48	60	72	80	90
PlayTennis:	No	No	Yes	Yes	Yes	No

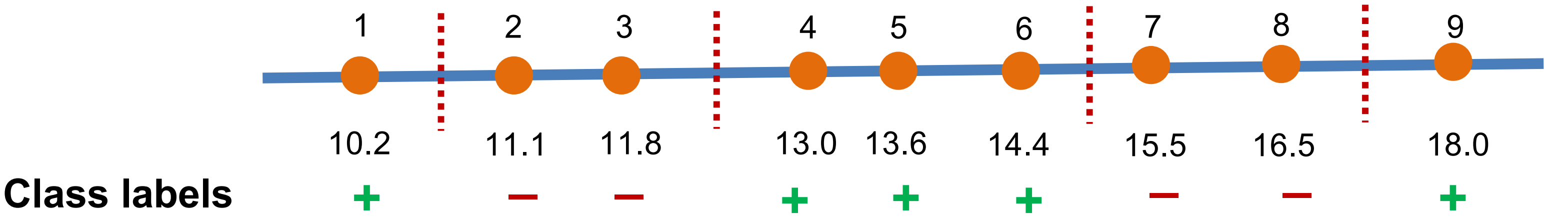
Continuous Valued Attributes

- Example for $n - 1$ splits



Continuous Valued Attributes

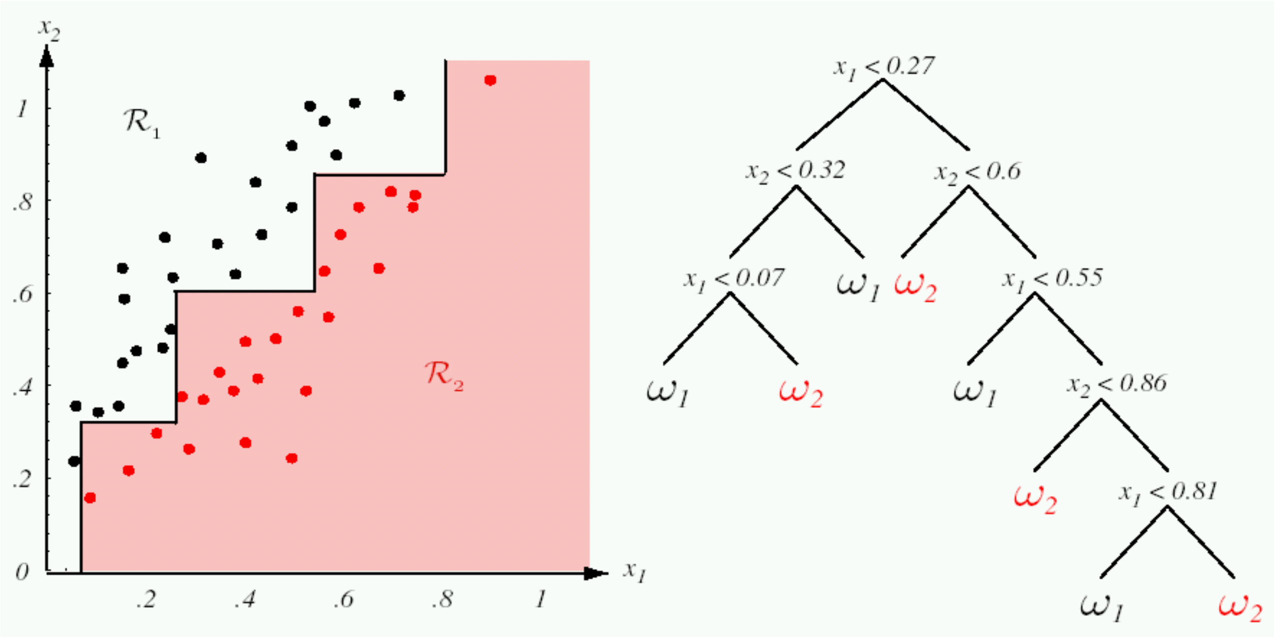
- Example for $n-1$ splits: more efficient number of splits if we know the class labels



- You can use the midway point as your boundary (e.g if $x > 12.4$ between points 3 and 4)

- Or, you can use the value in the training (e.g if $x > 11.8$ between points 3 and 4)

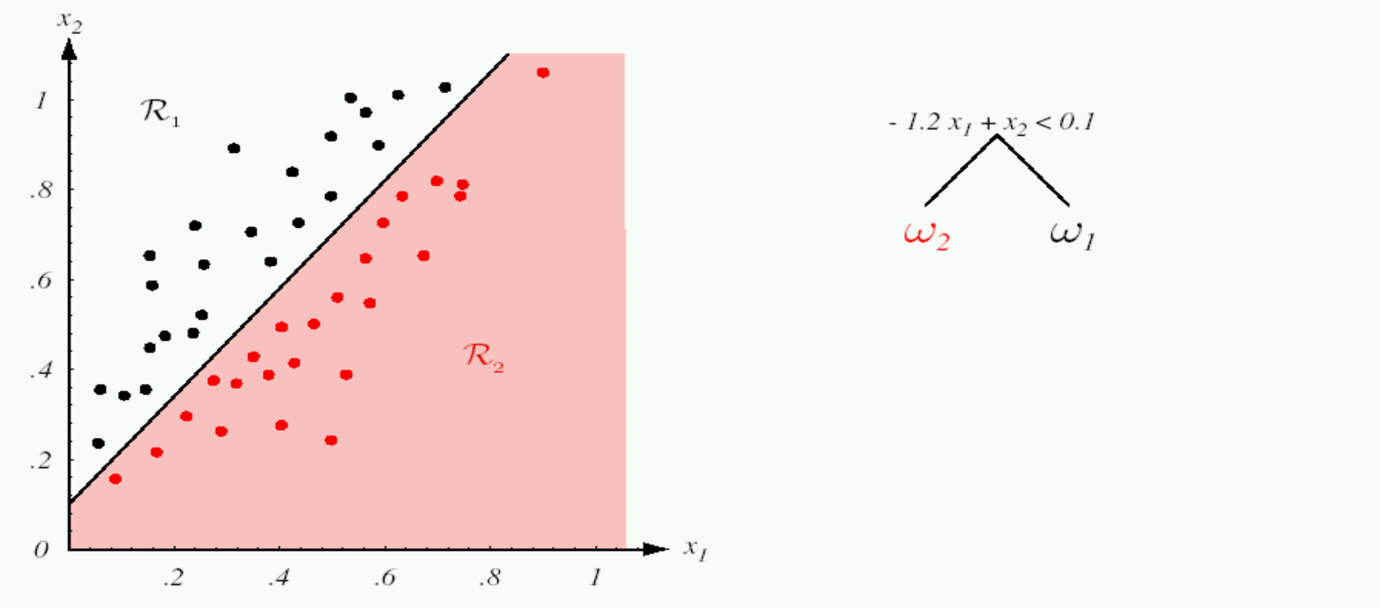
Axis-parallel Splitting



Fitting data that is not a good “match” to the possible splits in a tree.

“Pattern Classification” Duda, Hart, and Stork, (2001)

Splitting on Linear Combinations of Features



Reduced tree size by allowing splits that are a better “match” to the data.

“Pattern Classification” Duda, Hart, and Stork, (2001)

Attributes with Costs

Consider

- medical diagnosis, $Blood\ Test$ has cost $150
- robotics, computation cost 23 sec.

How to learn a consistent tree with low expected cost?

Attributes with Costs

One approach: evaluate information gain *relative* to cost:

- Example

$$\frac{Gain^2(S,A)}{Cost(A)}$$

Preference for decision trees using lower-cost attributes.

Misclassification with Costs

Also: class (misclassification) costs, instance costs, . . .

Can give *false positives* a different cost to *false negatives.*

Similar to expected loss in ”Classification 2” lecture, we can define a loss function $\lambda(\alpha_i|predicted\ class)$ (the loss associated to action α_i) and define the expected loss as:

$$R(\alpha_i|x) = \sum_{h \in H} \lambda(\alpha_i|predicted\ class)\ \ P(predicted\ class|x)$$

- however, if we use this loss function to split the tree, it has been shown that minimizing R does not necessarily lead to the best long-term growth of the tree.

Misclassification with Costs

So what is usually done instead :

- Grow the tree using gain information to its full depth (or other impurity measures)
- Apply cost minimization when post pruning (or minimize cost-complexity measure which takes the size of the tree into account as well)

Unknown Attribute Values

What if some examples missing values of A?

Use training example anyway, sort through tree. Here are 3 possible approaches

- If node n tests A, assign most common value of A among other examples sorted to node n
- assign most common value of A among other examples with same target value
- assign probability p_i to each possible value v_i of A
    - assign fraction p_i of example to each descendant in tree

Note: need to classify new (unseen) examples in same fashion

Windowing

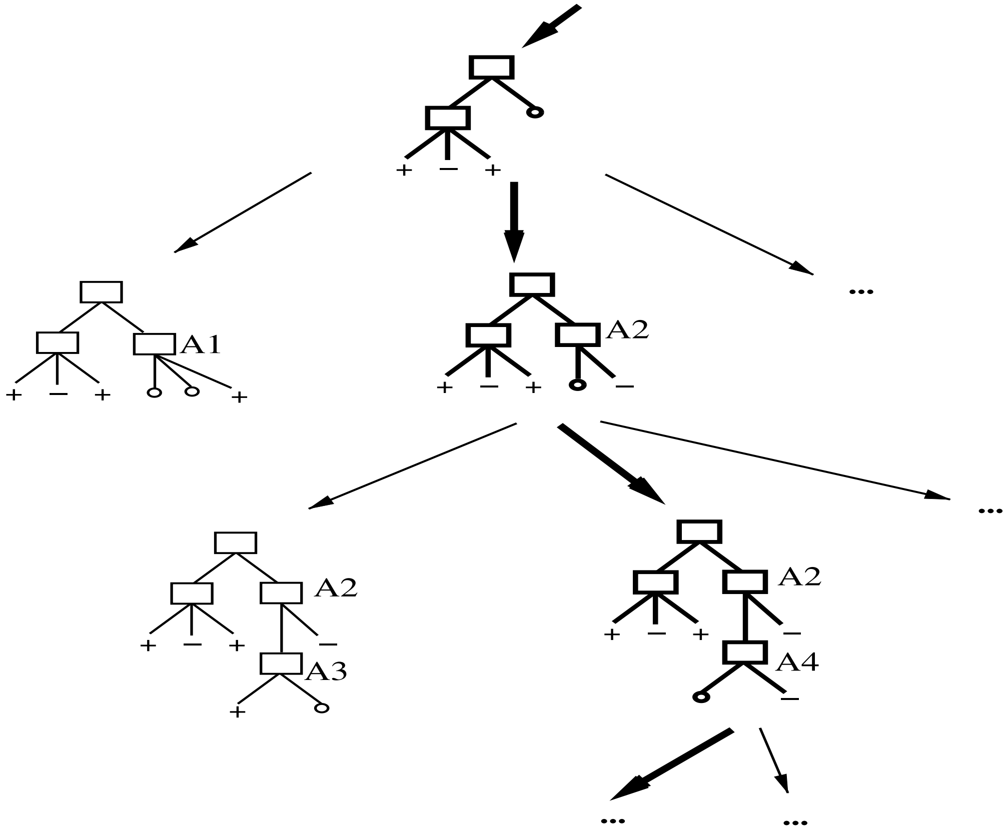
Early implementations – training sets was too large for memory. How to handle the problem?

As a solution ID3 implemented *windowing*:

1. select subset of instances – the *window*
2. construct decision tree from all instances in the window
3. use tree to classify training instances *not* in window
4. if all instances correctly classified then halt, else
5. add selected misclassified instances to the window
6. gotostep2

Windowing retained in C4.5 because it can lead to more accurate trees. Related to *ensemble learning*.

Hypothesis Space Search by ID3



Hypothesis Space Search by ID3

- Usual graph-search technique: greedy or beam search, starting with the node corresponding to the “empty tree” (single leaf node)
- Greedy choice: which one to select? The one that results in the greatest increase in posterior probability
- RESULT: set of trees with (reasonably) high posterior probabilities given D: we can now use these to answer questions like $P\ (y' \ = \ C_1 | \ldots)$? or even make a *decision* or a *classification* that $y' \ = \ C_1$, given input data x

Hypothesis Space Search by ID3

- ID3 searches the space of possible decision trees: doing hill-climbing on information gain.
- Hypothesis space is complete! (contains all finite discrete-valued functions w.r.t. attributes)
- Outputs a single hypothesis.
    - It cannot tell us how many other viable ones there are.
- No back tracking
    - Local minima...
- Uses all training examples at each step.
    - Results are less sensitive to noise
- Inductive bias

Inductive Bias in ID3

Inductive bias: set of assumptions needed in addition to training data to justify deductively learner's classifications

Restriction bias:

- The set of hypothesis that can be modelled by decision trees

Preference biases:

- Prefers trees with good splits near the top (splitting on features with the most information gain)
- Prefers shorter trees (comes naturally from good splits at the top and minimum description length)

Decision Tree

Pros:

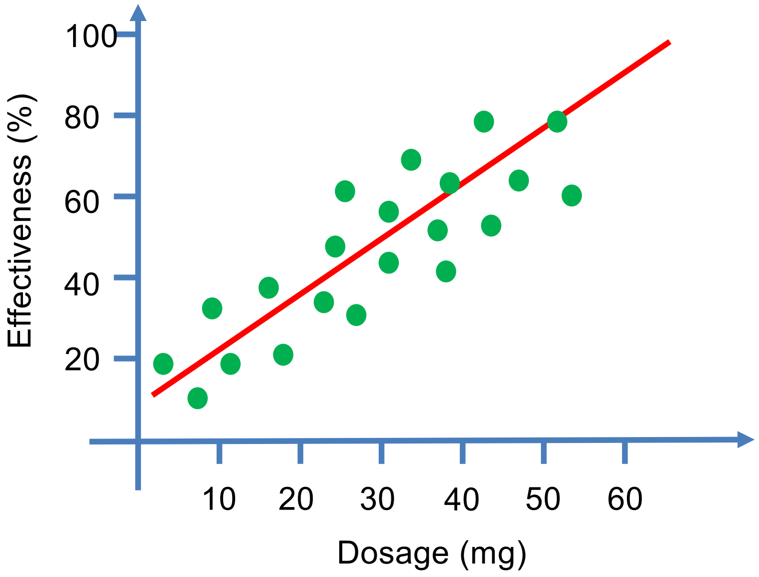
- Interpretability
- Easily handle irrelevant attributes (Gain = 0)
- Can handle both categorical and numerical data
- Can handle missing data
- Very compact (number of nodes << number of examples)
- Very fast at testing

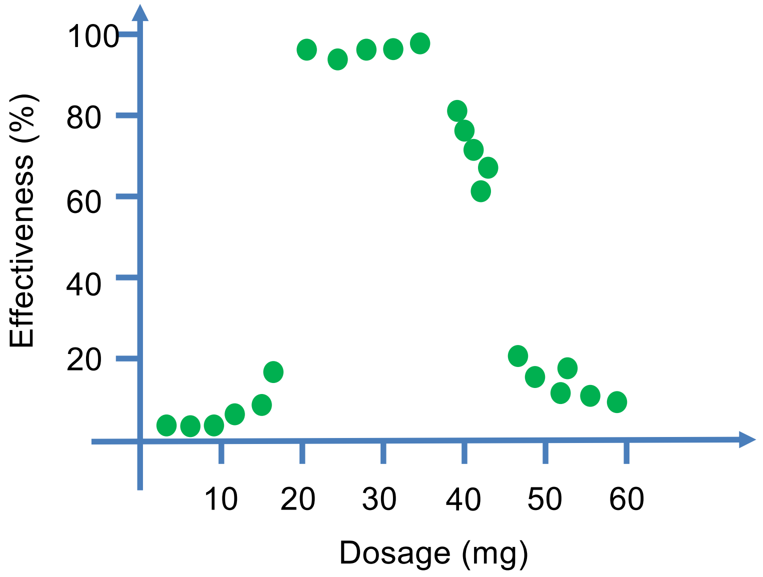
Cons:

- Only axis-aligned splits of the data
- Tend to overfit
- Greedy (may not find the best tree)
  - Exponentially many possible trees

Regression Tree

Example: A clinical trial has been done to evaluate the effect of a drug dosage.

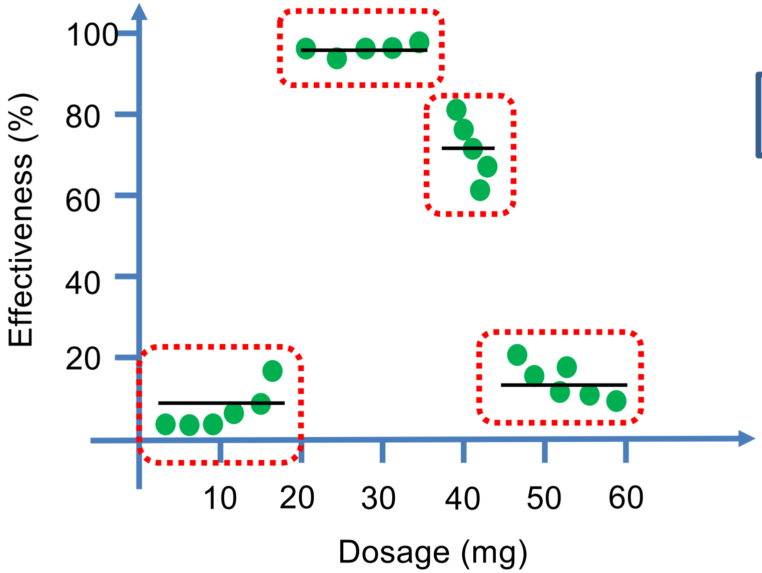


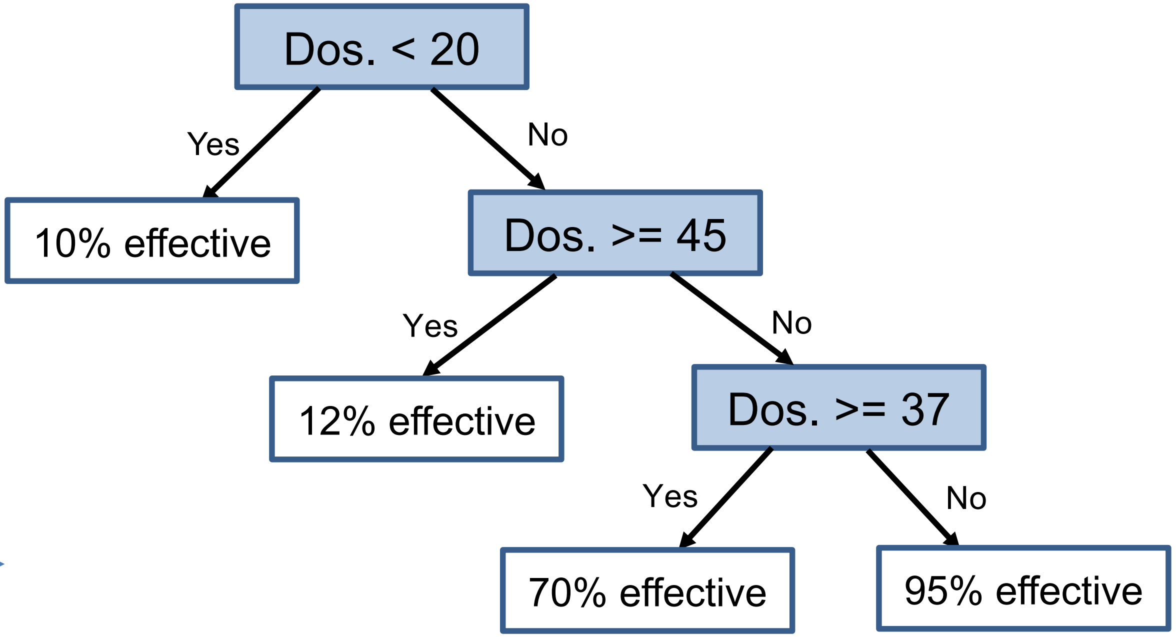


How to model the data on the right?

Regression Tree (CaRT algorithm)

One option would be to use regression trees





Each leaf corresponds to average drug effectiveness in a different cluster of examples, the tree does a better job than *Linear Regression*

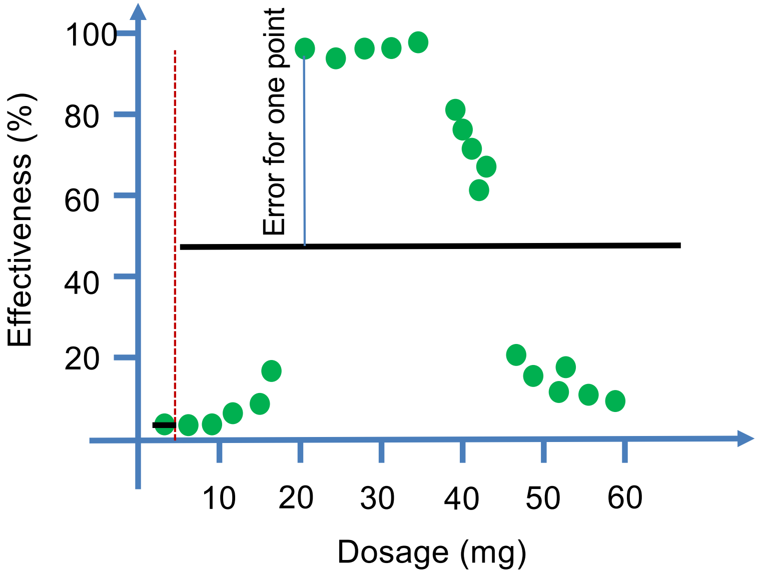
Regression Tree

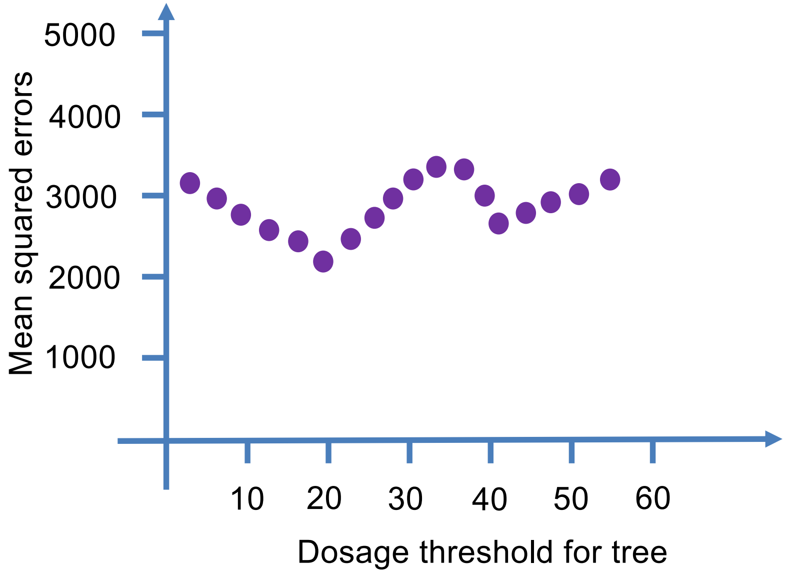
Although in this example, we can use other methods like *kNN* or *local regression*, but the advantage of *Regression Tree* will be more clear if you have multiple variables, specially if some of them are categorical and some are real-valued.

How to build the regression tree and find the thresholds?

Regression Tree

We will use mean squared errors to find the best threshold. We start from setting the threshold to the value of the first point to the last point, successively, and examine the mean of squared errors.





Regression Tree

- Mean squared error after setting the threshold, for each subset with m examples is:

$$MSE(Y_i) = \frac{1}{m}\sum_{j=1}^{m}(y_j - \bar{y})^2$$

  - The MSE here is equal to the variance of the examples in that subset

- Compute the weighted average of variance

$$weighted\ average\ variance = \sum_{i}^{l}\frac{|Y_i|}{|Y|}MSE(Y_i)$$

- We pick a threshold which minimizes the weighted-average/average of MSE/variance

Regression Tree

Multiple variables:

Imagine that in the drug experiment we also have other attributes like “age”, “sex”,…

To choose an attribute for each node:

- For each attribute separately find the best split using the minimum weighted average variance
- Pick the attribute with minimum weighted average variance

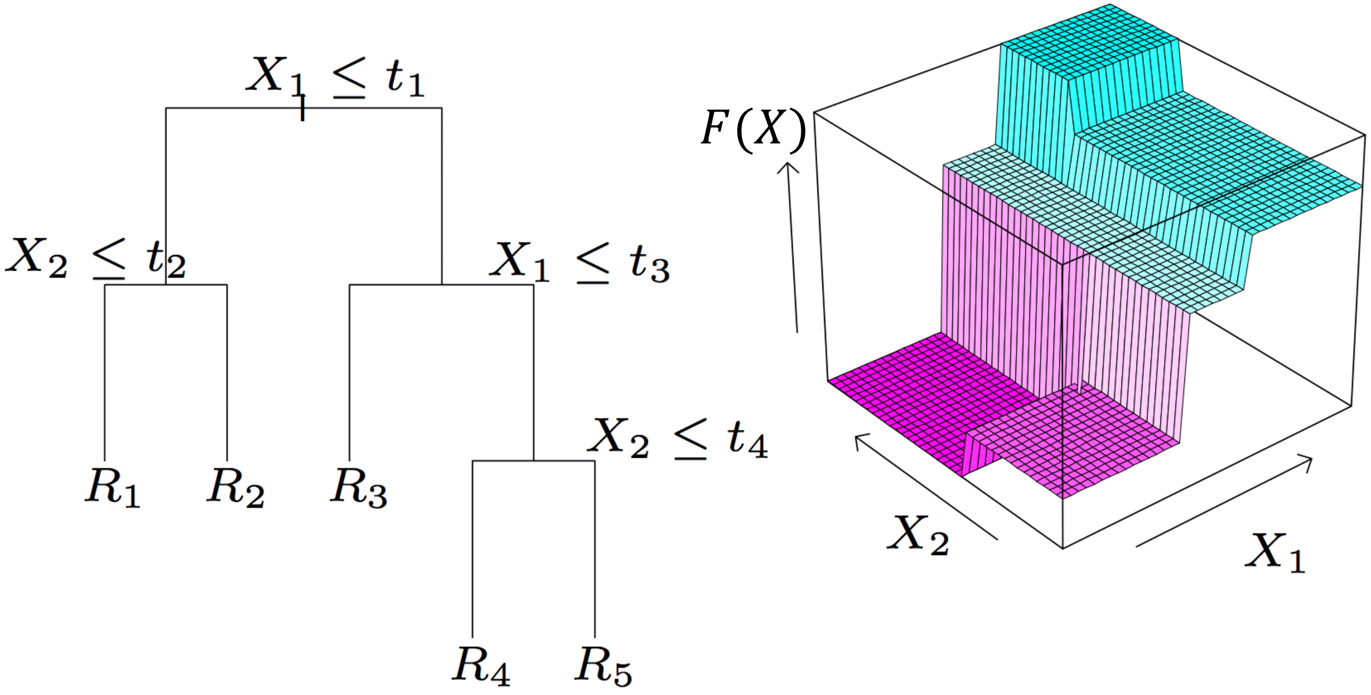
Regression Tree

Similar to decision tree, if we continue splitting all nodes that have a *mean squared error* bigger than zero, we will get a tree which fits the training data perfectly but will not generalize well to the test data.

How to avoid overfitting:

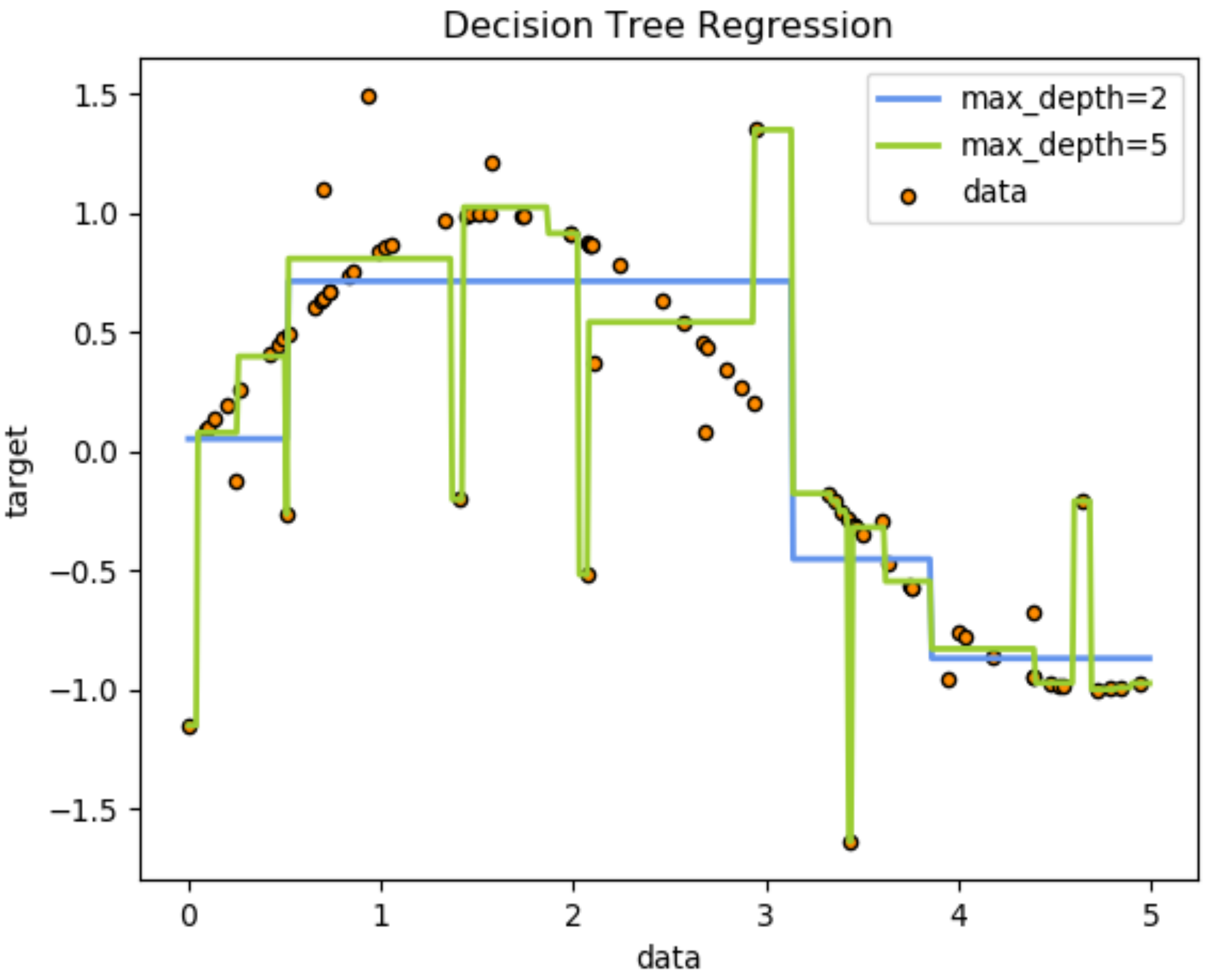
- The simplest to avoid overfitting, similar to, *decision tree* is to only split examples, when there are more than some minimum number (usually a default value for minimum number of examples to split is 20).
- Or using maximum depth. Allow the tree to grow to the specified depth
- Or maximum number of leaves
- …

A Regression Tree and its Prediction Surface

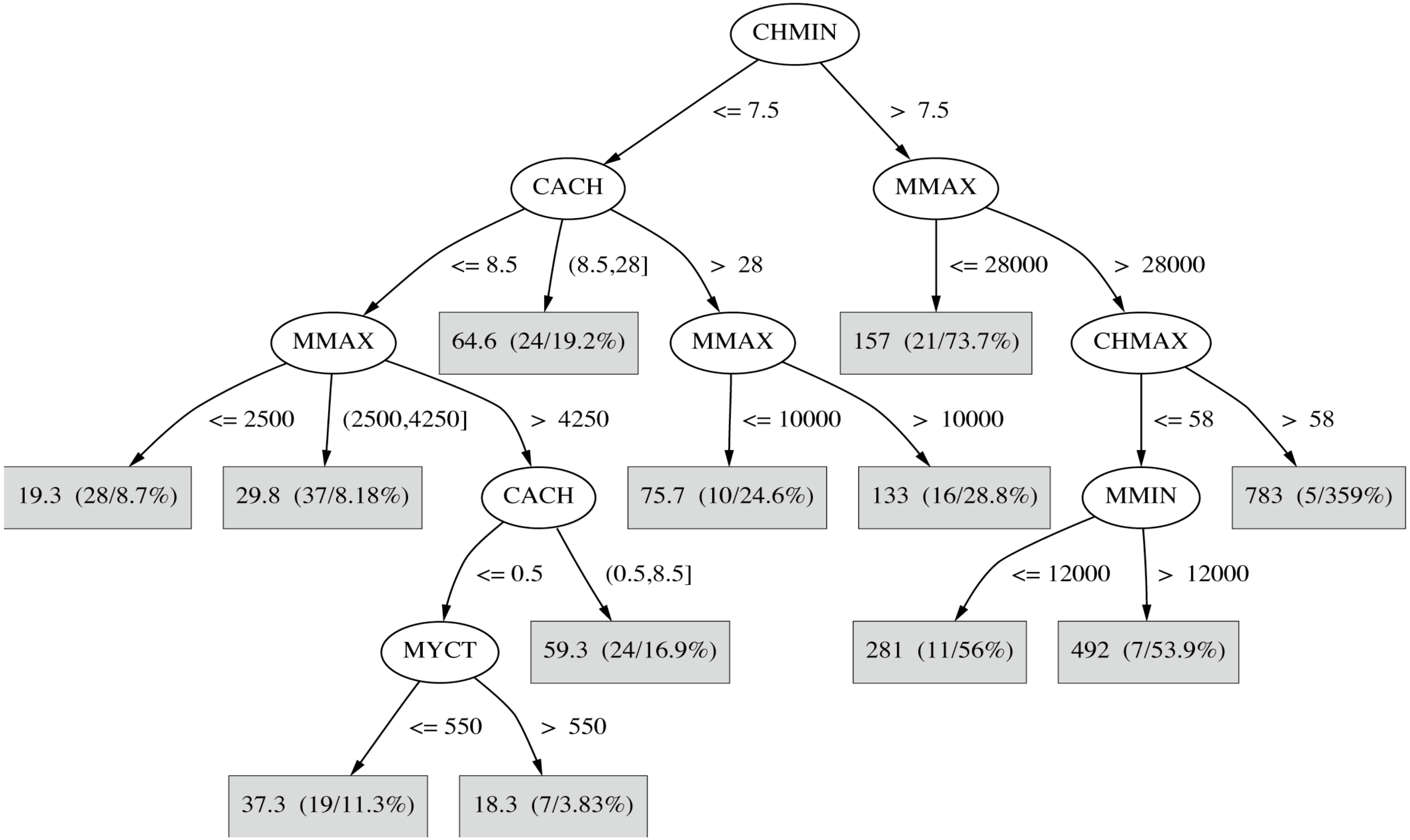


“Elements of Statistical Learning” Hastie, Tibshirani & Friedman (2001)

Regression Tree on sine dataset



Regression Tree on CPU dataset



Learning a regression tree

Imagine you are a collector of vintage Hammond tonewheel organs. You have been monitoring an online auction site, from which you collected some data about interesting transactions:

#	Model	Condition	Leslie	Price
1.	B3	excellent	no	4513
2.	T202	fair	yes	625
3.	A100	good	no	1051
4.	T202	good	no	270
5.	M102	good	yes	870
6.	A100	excellent	no	1770
7.	T202	fair	no	99
8.	A100	good	yes	1900
9.	E112	fair	no	77

Learning a regression tree

From this data, you want to construct a regression tree that will help you determine a reasonable price for your next purchase.

There are three features, hence three possible splits:

$$Model \;=\; [A100, B3, E112, M102, T202]$$
$$[1051, 1770, 1900][4513][77][870][99, 270, 625]$$

$$Condition \;=\; [excellent, good, fair]$$
$$[1770, 4513][270, 870, 1051, 1900][77, 99, 625]$$

$$Leslie = [yes, no]$$
$$[625{,}870{,}1900][77{,}99{,}270{,}1051{,}1770{,}4513]$$

Learning a regression tree

- The means of the first split are 1574, 4513, 77, 870 and 331, and the weighted average of mean squared errors is 6.25×10^4.
- The means of the second split are 3142, 1023 and 267, with weighted average of mean squared errors 5.9×10^5;
- for the third split the means are 1132 and 1297, with weighted average of mean squared errors 1.72×10^6.

We therefore branch on $Model$ at the top level. This gives us three single-instance leaves, as well as three $A100$s and three $T202$s.

Learning a regression tree

For the A100s we obtain the following splits:

$$Condition = [excellent, good, fair]$$
$$[1770][1051{,}1900][\quad]$$

$$Leslie = [yes, no]$$
$$[1900][1051{,}1770]$$

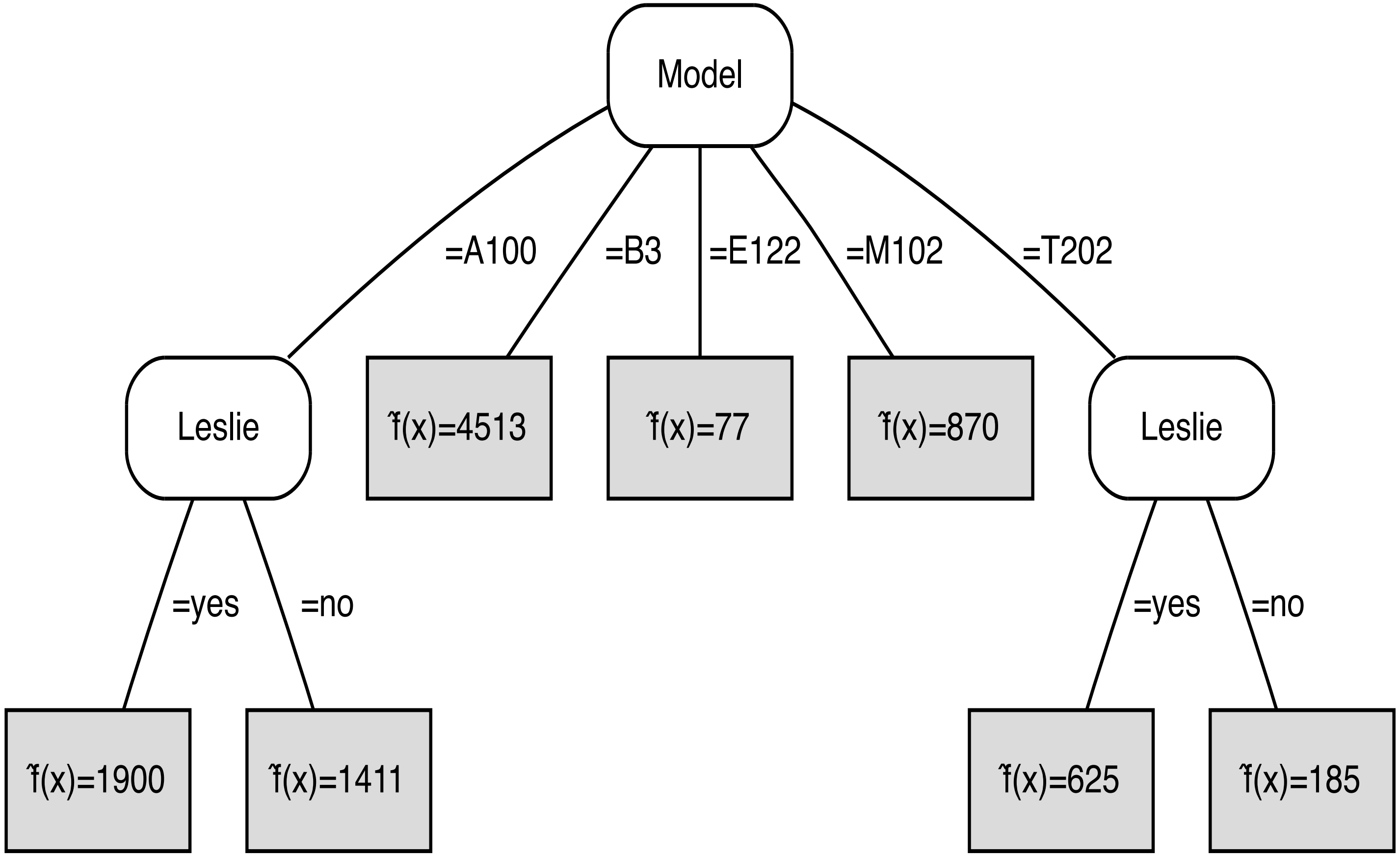
Without going through the calculations, we can see that the second split results in less variance (to handle the empty child, it is customary to set its variance equal to that of the parent). For the $T202$s the splits are as follows:

$$Condition = [excellent, good, fair]$$
$$[][270][99{,}625]$$

$$Leslie = [yes, no]$$
$$[625][99{,}270]$$

Again we see that splitting on Leslie gives tighter clusters of values. The learned regression tree is depicted on the next slide.

A regression tree



A regression tree learned from the Hammond organ dataset.

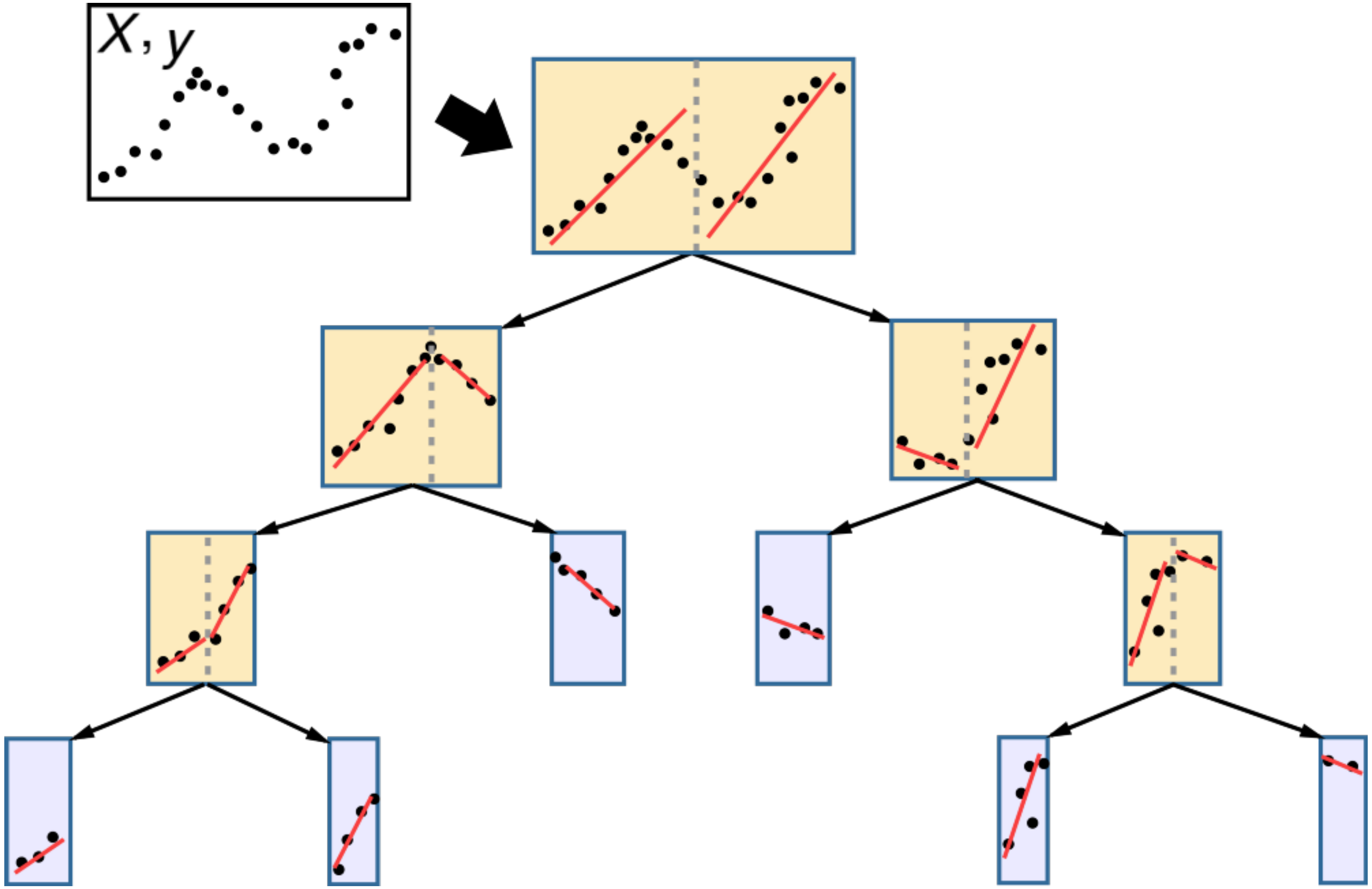
Regression trees

- Differences to decision trees:
  - Splitting criterion: minimizing intra-subset variation
  - Pruning criterion: based on numeric error measure
  - Leaf node predicts average class values of training instances reaching that node
- Can approximate piecewise constant functions
- Easy to interpret
- More sophisticated version: model trees

Model trees

- Like regression trees but with linear regression functions (or any model of your choice) at each node
- Linear regression applied to instances that reach a node after full tree has been built
- Only a subset of the attributes is used for Linear Regression
  - Attributes occurring in subtree (+maybe attributes occurring in path to the root)
- Fast: overhead for Linear Regression not large because usually only a small subset of attributes is used in tree

Model trees

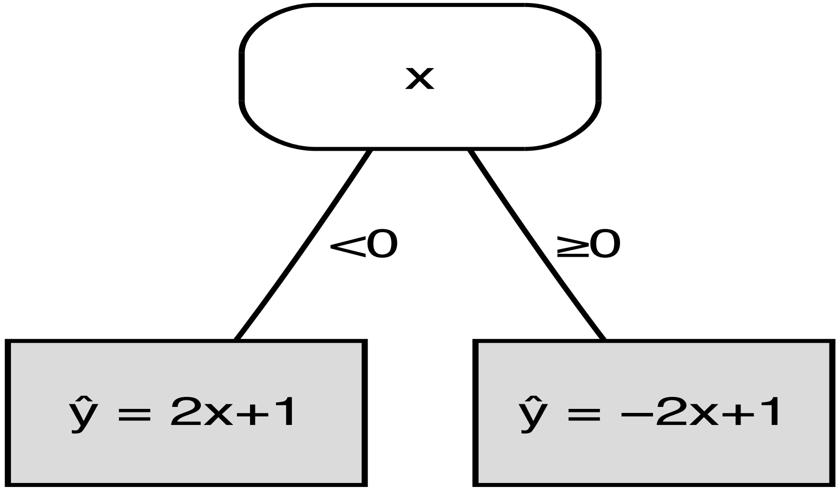


Model Tree Example

Suppose we want to approximate $y = \cos \pi x$ on the interval $-1 \leq x \leq 1$.

- A linear approximation is not much use here, since the best fit would be $y = 0$.
- However, if we split the x-axis in two intervals $-1 \leq x < 0$ and $0 \leq x \leq 1$, we could find reasonable linear approximations on each interval. We can achieve this by using x both as a splitting feature and as a regression variable (next slide).

Model Tree Example



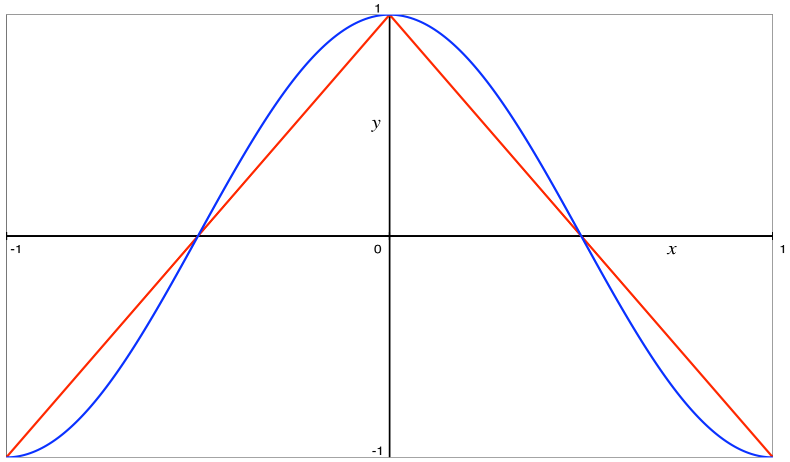
x

<0

≥ 0

$\hat{y} = 2x+1$

$\hat{y} = -2x+1$



Building the tree

- Splitting criterion: *standard deviation reduction*

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

where $T_1,\ T_2, \ldots$ are the sets from splits of data at node.

- Termination criteria (important when building trees for numeric prediction):
  - Standard deviation becomes smaller than certain fraction of sd for full training set (e.g. 5%)
  - Too few instances remain (e.g. less than four)

Pruning the tree

Model trees suffer the same deficits as decision tree, which is the tendency to overfit the train data when using complex models

- Pruning is based on estimated absolute error of LR models
- Linear Regression (LR) models are pruned by greedily removing terms to minimize the estimated error
- Model trees allow for heavy pruning: often a single LR model can replace a whole subtree
- Pruning proceeds bottom up: error for LR model at internal node is compared to error for subtree

Discrete (nominal) attributes

- Nominal attributes converted to binary attributes and treated as numeric
  - Nominal values sorted using average class value for each one
  - For k-values, k binary attributes are generated
    - the ith binary attribute is 0 if an instance's value is one of the first i in the ordering, 1 otherwise
- Best binary split with original attribute provably equivalent to a split on one of the new attributes

Summary – decision trees

- Decision tree learning is a practical method for many classifier learning tasks – still a "Top 10" data mining algorithm
- TDIDT family descended from ID3 searches complete hypothesis space - the hypothesis is there, somewhere...
- Overfitting is inevitable with an expressive hypothesis space and noisy data, so pruning is important
- Decades of research into extensions and refinements of the general approach, e.g., for numerical prediction, logical trees

Summary – regression and model trees

- Often the “try-first” machine learning method in applications, illustrates many general issues
- Performance can be improved with use of “ensemble” methods
- Regression trees were introduced in CaRT – R’s implementation is close to CaRT, but see sklearn.tree.DecisionTreeRegressor for a basic version

Acknowledgements

- Material derived from slides for the book “Machine Learning” by T. Mitchell McGraw-Hill (1997) http://www-2.cs.cmu.edu/~tom/mlbook.html

- Material derived from slides by Andrew W. Moore

- http:www.cs.cmu.edu/~awm/tutorials Material derived from slides by Eibe Frank

- http://www.cs.waikato.ac.nz/ml/weka Material derived from slides for the book “Machine Learning” by P. Flach Cambridge University Press (2012) http://cs.bris.ac.uk/~flach/mlbook

- Material derived from slides by Josh Starmer (StatQuest Channel)