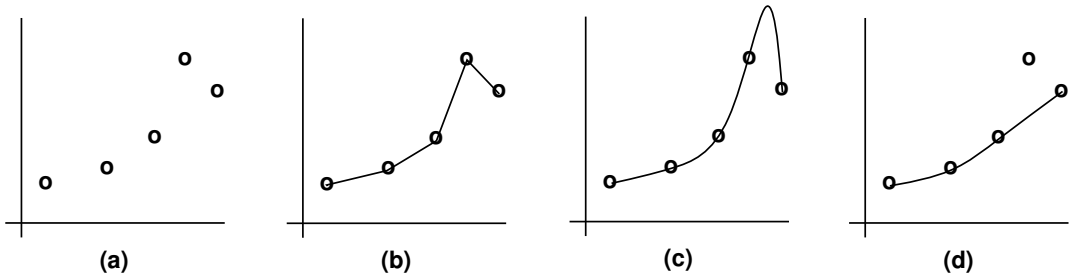
Decision Tree Learning

- Blue slides: Mitchell
- Olive slides: Alpaydin

Inductive Learning (Refresher)

- Given **example** pairs $(x, f(x))$, return a function h (the hypothesis) that approximates the function f:
  - **pure inductive inference,** or **induction**.
  - Supervised learning is basically a form of inductive learning.

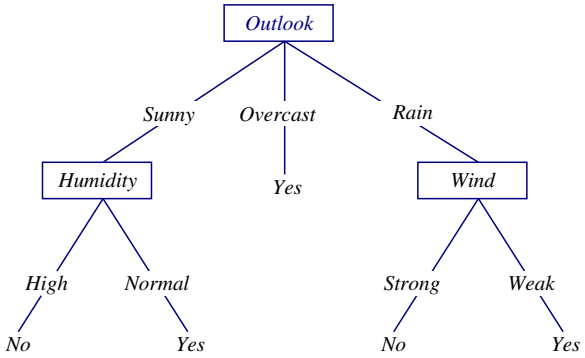
Inductive Bias



Given (a) as the training data, we can come up with several different hypotheses: (b) to (d)

- selection of one hypothesis over another is called a **bias**.
  - exact match to training data
  - prefer imprecise but smooth approximation
  - prefer simpler models with fewer parameters
  - etc.

Decision Tree Learning



- Learn to approximate **discrete-valued** target functions.
- Step-by-step decision making: It can learn **disjunctive expressions**: Hypothesis space is **completely expressive**, avoiding problems with restricted hypothesis spaces.
- Inductive bias: **small trees** over large trees.

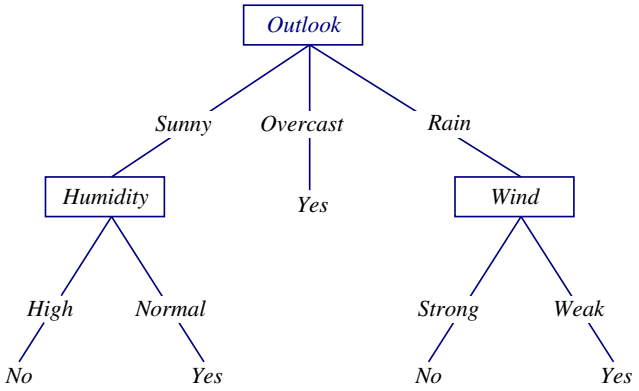
Example

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

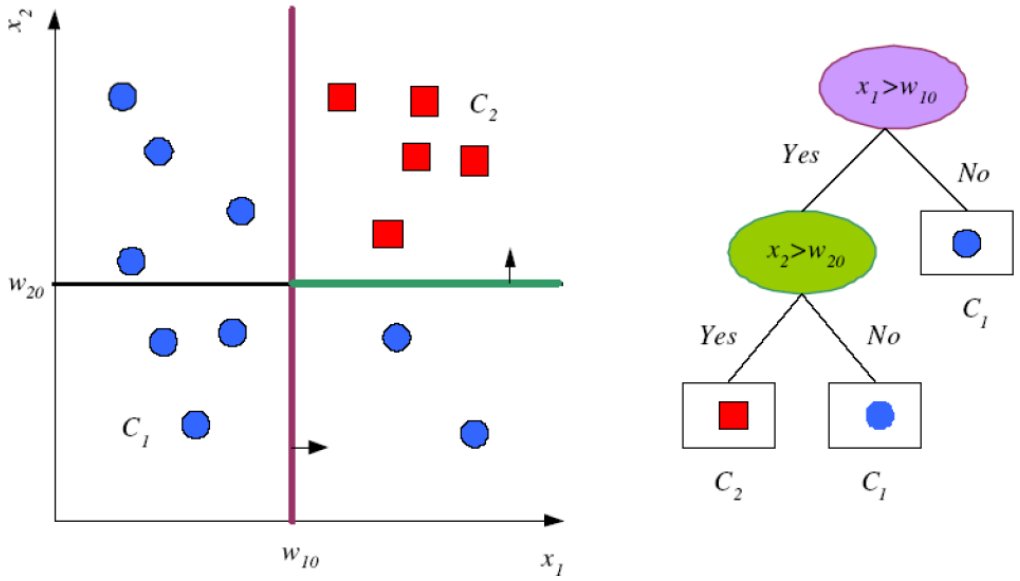
- A popular **inductive inference** algorithm.
- Algorithms: ID3, ASSISTANT, C4.5, etc.
- Applications: medical diagnosis, assess credit risk of loan applicants, etc.

Decision Trees: Operation



- Each instance holds attribute values.
- Instances are classified by filtering the attribute values down the decision tree, down to a leaf which gives the final answer.
- Internal nodes: attribute names or attribute values. Branching occurs at attribute nodes.

Tree Uses Nodes and Leaves (ALP)

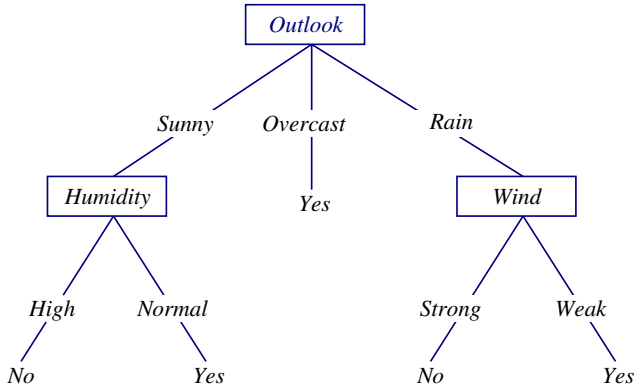


- Each node forms a decision boundary.
- Each leaf represents a class (or numerical value)

Divide and Conquer (ALP)

- Internal decision nodes
  - Univariate: Uses a single attribute, x_i
    - Numeric x_i: Binary split based on comparison $x_i > w_m$
    - Discrete x_i: n-way split for n possible values
  - Multivariate: Uses all attributes, $\vec{x}$
- Leaves
  - Classification: Class labels
  - Regression: Numeric value
- Learning is greedy: Find the best split recursively. (Breiman et al., 1984; Quinlan, 1986, 1993)

Decision Trees: What They Represent



- Each path from root to leaf is a **conjunctions of constraints** on the attribute values.

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

Appropriate Tasks for Decision Trees

Good at **classification problems** where:

- Instances are represented by **attribute-value pairs**.
- The target function has **discrete output values**.
- Disjunctive descriptions may be required.
- The training data **may contain errors**.
- The training data **may contain missing attribute values**.

Constructing Decision Trees from Examples

- Given a set of examples (**training set**), both **positive** and **negative**, the task is to construct a decision tree that describes a concise decision path.
- Using the resulting decision tree, we want to **classify** new instances of examples (either as **yes** or **no**).

Constructing Decision Trees: Trivial Solution

- A trivial solution is to explicitly construct paths for each given example. In this case, you will get a tree where the number of leaves is the same as the number of training examples.
- The problem with this approach is that it is not able to deal with situations where, some attribute values are missing or new kinds of situations arise.
- Consider that some attributes may not count much toward the final classification.

Finding a Concise Decision Tree

- Memorizing all cases may not be the best way.
- We want to extract a decision pattern that can describe a large number of cases in a **concise** way.
- In terms of a decision tree, we want to make as few tests as possible before reaching a decision, i.e. the depth of the tree should be shallow.

Finding a Concise Decision Tree (cont'd)

- Basic idea: pick up attributes that can clearly separate positive and negative cases.
- These attributes are more important than others: the final classification heavily depend on the value of these attributes.

Decision Tree Learning Algorithm: ID3

Main loop:

1. $A \leftarrow$ the "best" decision attribute for next $node$
2. Assign A as decision attribute for $node$
3. For each value of A, create new descendant of $node$
4. Sort training examples to leaf nodes
5. If training examples perfectly classified, Then STOP, Else iterate over new leaf nodes

ID3 stands for Iterative Dichotomizer 3

Choosing the Best Attribute

$A1$ or $A2$?

- With initial and final number of positive and negative examples based on the attribute just tested, we want to decide which attribute is **better**.
- How to quantitatively measure which one is better?

Choosing the Best Attribute to Test First

Use Shannon's information theory to choose the attribute that give the maximum **information gain**.

- Pick an attribute such that the information gain (or entropy reduction) is maximized.
- Entropy measures the **average surprisal** of events. Less probable events are more surprising.

Information Theory (Informal Intro)

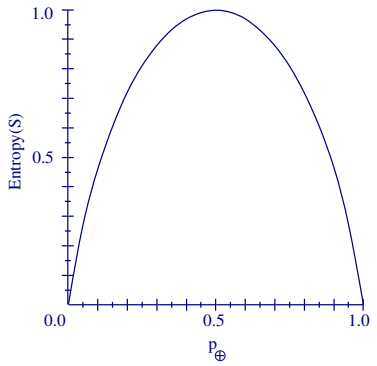
Given two events, H and T (Head and Tail):

- Rare (uncertain) events give more surprise:
  H more surprising than T if $P(H) < P(T)$
  H more uncertain than T if $P(H) < P(T)$
- How to represent "more surprising", or "more uncertain"?
  $Surprise(H) > Surprise(T)$ if

$$
\begin{aligned}
& P(H) < P(T) \\
\leftrightarrow \quad & \frac{1}{P(H)} > \frac{1}{P(T)} \\
\leftrightarrow \quad & \log\left(\frac{1}{P(H)}\right) > \log\left(\frac{1}{P(T)}\right) \\
\leftrightarrow \quad & -\log\left(P(H)\right) > -\log\left(P(T)\right)
\end{aligned}
$$

- $-\log\left(P(X)\right)$ as a measure of **uncertainty**.

Information Theory (Cont'd)



- 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
- Entropy measures the average uncertainty in S

$$Entropy(S) \equiv -p_\oplus \log_2 p_\oplus - p_\ominus \log_2 p_\ominus$$

Uncertainty and Information

- By performing some query, if you go from state S_1 with entropy $E(S_1)$ to state S_2 with entropy $E(S_2)$, where $E(S_1) > E(S_2)$, your **uncertainty has decreased**.
- The amount by which uncertainty decreased, i.e., $E(S_1) - E(S_2)$, can be thought of as information you gained (**information gain**) through getting answers to your query.

Entropy and Code Length

- $Entropy(S)$ = expected number of bits needed to encode class ($\oplus$ or $\ominus$) of randomly drawn member of S (under the optimal, shortest-length code)
- Information theory: optimal length code assigns $-\log_2 p$ bits to message having probability p.
  - Encode with short string for frequent messages (less surprising), and long string for rarely occurring messages (more surprising).
- So, expected number of bits to encode $\oplus$ or $\ominus$ of random member of S:

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

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

Entropy and Information Gain

$$Entropy(S) = \sum_{i \in C} -P_i log_2(P_i)$$

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

- C: categories (classifications)
- S: set of examples
- A: a single attribute
- S_v: set of examples where attribute $A = v$.
- $|X|$: cardinality of arbitrary set X.

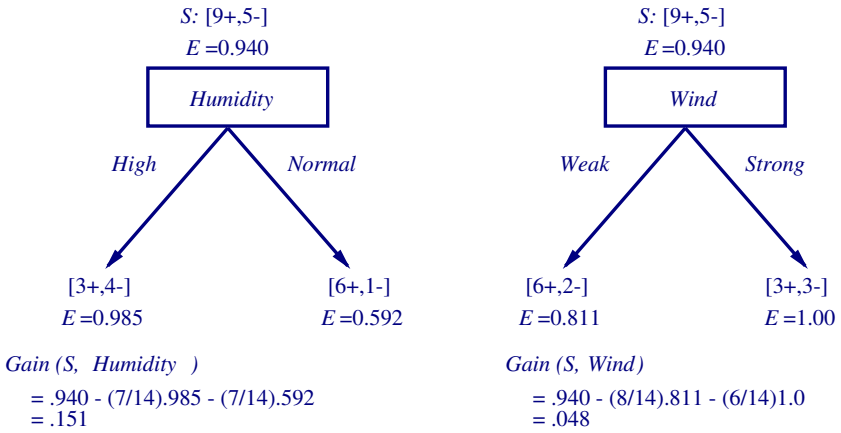
Example

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

- Which attribute to test first?

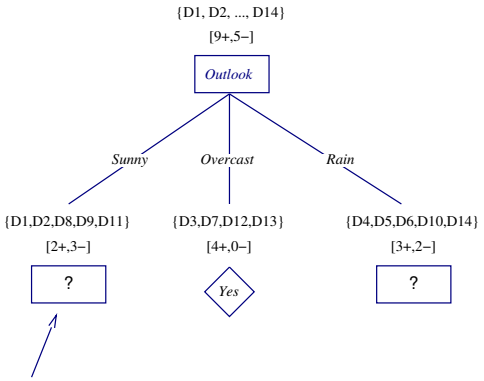
Choosing the Best Attribute

Which attribute is the best classifier?



- $+$: # of positive examples; $-$: # of negative examples
- Initial entropy $= -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14} = 0.94$.
- You can calculate the rest.
- Note: $0.0 \times \log 0.0 \equiv 0.0$ even though $\log 0.0$ is not defined.

Partially Learned Tree



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

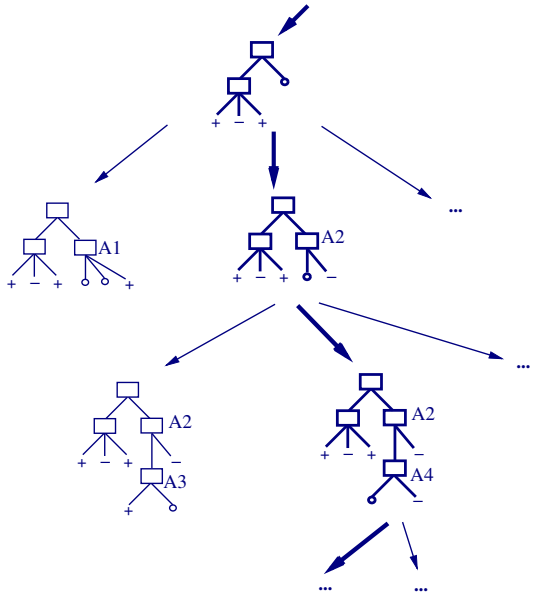
Gain (S_{sunny}, *Humidity*) = .970 − (3/5) 0.0 − (2/5) 0.0 = .970

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

- Select next attribute, based on the remaining examples.

Hypothesis Space Search in ID3



- At each branch, we make a decision regarding a particular attribute. Choice of an attribute directs the search toward a certain final hypothesis.

Hypothesis Space Search in ID3

- Hypothesis space is complete!
  - Target function surely in there...
- Outputs a single hypothesis (which one?)
  - Can't play 20 questions...
- No back tracking
  - Local minima...
- Statistically-based search choices
  - Robust to noisy data...
- Inductive bias: approx "prefer shortest tree"

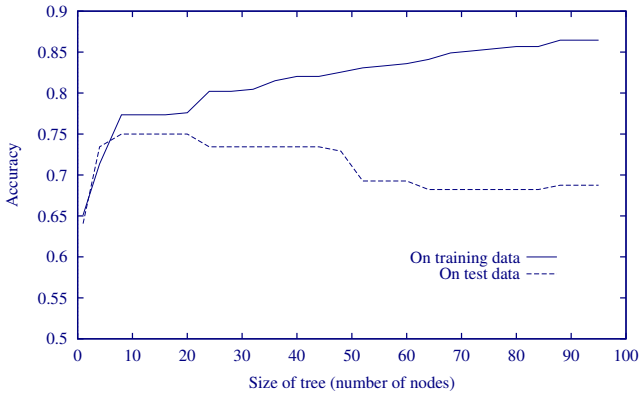
Inductive Bias in ID3

- ID3 is biased:
  - Not because of the restriction on the hypothesis space, but
  - Because of the **preference** for a particular hypothesis.
- Such an inductive bias is called **Occam's razor**: *The most likely hypothesis is the* **simplest one** *that is consistent with all observations.*

Accuracy of Decision Trees

- Divide examples into training and test sets.
- Train using the training set.
- Measure accuracy of resulting decision tree on the test set.

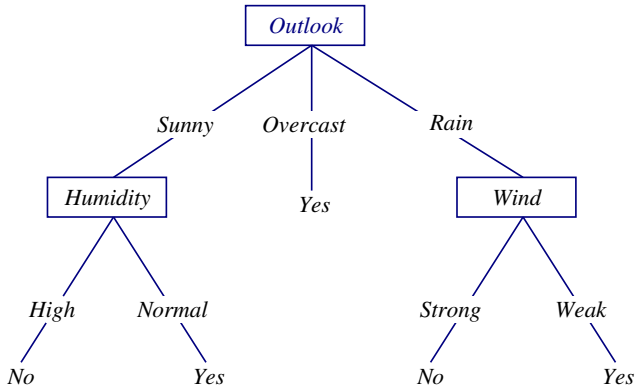
Issue: Overfitting



Overfitting:

- Given a hypothesis space H, a hypothesis $h \in H$ is said to **overfit** the training data if there exists some alternative hypothesis $h' \in H$ such that h' is worse than h on the training set but h' is better than h over the entire distribution of instances.
- Can be due to **noise** in data.

Issue: Noise



- What if $\langle Outlook = Sunny, Temp = Hot, Humidity = Normal, Wind = Strong, Play = No \rangle$ was added as a training example?
- Further elaboration of the above tree becomes necessary.
- The resulting tree will fit the training data plus the noise, but it may perform poorly on the true instance distribution.

Overcoming Overfitting

- Stop early.
- Allow overfitting, then post-prune tree.
- Use separate set of examples not used in training to monitor performance on unobserved data (validation set).
- Use all available data, but perform statistical test to estimate chance of improving.
- Use explicit measure of complexity of encoding, and put a bound on tree size.

Regression Trees (ALP)

- Error at node m:

$$b_m(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{X}_m : \mathbf{x} \text{ reaches node } m \\ 0 & \text{otherwise} \end{cases}$$

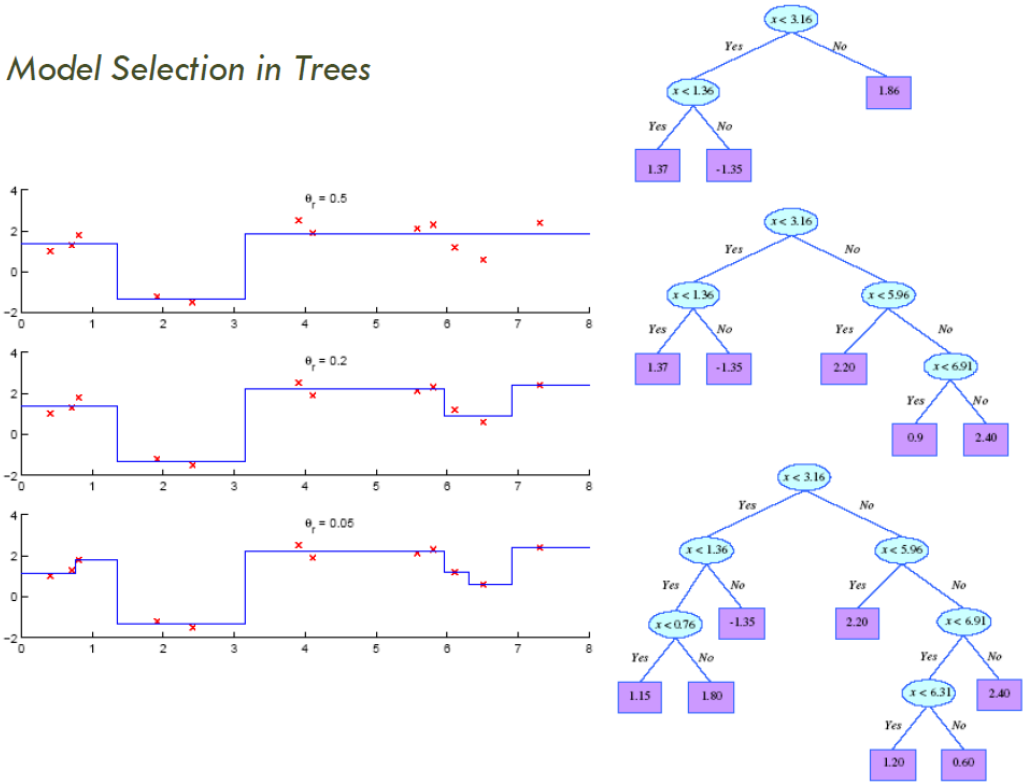
$$E_m = \frac{1}{N_m} \sum_t \left(r^t - g_m\right)^2 b_m\left(\mathbf{x}^t\right) \qquad g_m = \frac{\sum_t b_m\left(\mathbf{x}^t\right) r^t}{\sum_t b_m\left(\mathbf{x}^t\right)}$$

- After splitting:

$$b_{mj}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{X}_{mj} : \mathbf{x} \text{ reaches node } m \text{ and branch } j \\ 0 & \text{otherwise} \end{cases}$$

$$E'_m = \frac{1}{N_m} \sum_j \sum_t \left(r^t - g_{mj}\right)^2 b_{mj}\left(\mathbf{x}^t\right) \qquad g_{mj} = \frac{\sum_t b_{mj}\left(\mathbf{x}^t\right) r^t}{\sum_t b_{mj}\left(\mathbf{x}^t\right)}$$

Model Selection in Regression Trees (ALP)



Pruning Trees (ALP)

- Remove subtrees for better generalization (decrease variance)
  - Prepruning: early stopping
  - Postpruning: grow the whole tree, then prune subtree that overfits on the pruning set.
- Prepruning is fater, postpruning is more accurate (but requires separate pruning set)

Other Issues

- Rule extraction ; Rule leanring
- Continuous-valued attributes: dynamically define new discrete-valued attributes
- Multi-valued attributes with large number of possible values: Use measures other than information gain.
- Training examples with missing attribute values: Assign most common value, or assign with the occurring frequency.
- Attributes with different cost/weighting: Scale using the cost.