celeb	yes
Math'n	NO
Male	Y
Political	N
USA	Y
Real	Y
ACTOR	N
ALIVE	N
M Jackson	N

MUSIC	N
ACAD	N
COMED	N
AUTHOR	N
SCIENCE	Y
STEVE JOBS	Y

Query Selection

- Key Question: **Given a partial tree down to node N, what feature s should we choose for the property test T?**
- The obvious heuristic is to choose the feature that yields as big a decrease in the impurity as possible.
- The impurity gradient is

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) , \tag{5}$$

where N_L and N_R are the left and right descendants, respectively, P_L is the fraction of data that will go to the left sub-tree when property T is used.

Query Selection

- Key Question: **Given a partial tree down to node N, what feature s should we choose for the property test T?**
- The obvious heuristic is to choose the feature that yields as big a decrease in the impurity as possible.
- The impurity gradient is

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) \ , \tag{5}$$

  where N_L and N_R are the left and right descendants, respectively, P_L is the fraction of data that will go to the left sub-tree when property T is used.
- The strategy is then to choose the feature that maximizes $\Delta i(N)$.

Query Selection

- Key Question: **Given a partial tree down to node N, what feature s should we choose for the property test T?**
- The obvious heuristic is to choose the feature that yields as big a decrease in the impurity as possible.

COVER & Thomas

- The impurity gradient is

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) \quad , \tag{5}$$

  where N_L and N_R are the left and right descendants, respectively, P_L is the fraction of data that will go to the left sub-tree when property T is used.
- The strategy is then to choose the feature that maximizes $\Delta i(N)$.
- If the **entropy impurity** is used, this corresponds to choosing the feature that yields the highest **information gain**.

What can we say about this strategy?

- For the binary-case, it yields one-dimensional optimization problem (which may have non-unique optima).

What can we say about this strategy?

- For the binary-case, it yields one-dimensional optimization problem (which may have non-unique optima).
- In the higher branching factor case, it would yield a higher-dimensional optimization problem.
  - In multi-class binary tree creation, we would want to use the **twoing criterion**. The goal is to find the split that best separates groups of the c categories. A candidate "supercategory" C_1 consists of all patterns in some subset of the categories and C_2 has the remainder. When searching for the feature s, we also need to search over possible category groupings.

What can we say about this strategy?

- For the binary-case, it yields one-dimensional optimization problem (which may have non-unique optima).
- In the higher branching factor case, it would yield a higher-dimensional optimization problem.
  - In multi-class binary tree creation, we would want to use the **twoing criterion**. The goal is to find the split that best separates groups of the c categories. A candidate "supercategory" C_1 consists of all patterns in some subset of the categories and C_2 has the remainder. When searching for the feature s, we also need to search over possible category groupings.
- This is a local, greedy optimization strategy.

What can we say about this strategy?

- For the binary-case, it yields one-dimensional optimization problem (which may have non-unique optima).
- In the higher branching factor case, it would yield a higher-dimensional optimization problem.
    - In multi-class binary tree creation, we would want to use the **twoing criterion**. The goal is to find the split that best separates groups of the c categories. A candidate "supercategory" C_1 consists of all patterns in some subset of the categories and C_2 has the remainder. When searching for the feature s, we also need to search over possible category groupings.
- This is a local, greedy optimization strategy.
- Hence, there is no guarantee that we have either the global optimum (in classification accuracy) or the smallest tree.

What can we say about this strategy?

- For the binary-case, it yields one-dimensional optimization problem (which may have non-unique optima).
- In the higher branching factor case, it would yield a higher-dimensional optimization problem.
  - In multi-class binary tree creation, we would want to use the **twoing criterion**. The goal is to find the split that best separates groups of the c categories. A candidate "supercategory" C_1 consists of all patterns in some subset of the categories and C_2 has the remainder. When searching for the feature s, we also need to search over possible category groupings.
- This is a local, greedy optimization strategy.
- Hence, there is no guarantee that we have either the global optimum (in classification accuracy) or the smallest tree.
- In practice, it has been observed that the particular choice of impurity function rarely affects the final classifier and its accuracy.

A Note About Multiway Splits

- In the case of selecting a multiway split with branching factor B, the following is the direct generalization of the impurity gradient function:

$$\Delta i(s) = i(N) - \sum_{k=1}^{B} P_k i(N_k) \tag{6}$$

A Note About Multiway Splits

- In the case of selecting a multiway split with branching factor B, the following is the direct generalization of the impurity gradient function:

$$\Delta i(s) = i(N) - \sum_{k=1}^{B} P_k i(N_k) \tag{6}$$

- This direct generalization is biased toward higher branching factors.
  - To see this, consider the uniform splitting case.

A Note About Multiway Splits

- In the case of selecting a multiway split with branching factor B, the following is the direct generalization of the impurity gradient function:

$$\Delta i(s) = i(N) - \sum_{k=1}^{B} P_k i(N_k) \tag{6}$$

- This direct generalization is biased toward higher branching factors.
  - To see this, consider the uniform splitting case.
- So, we need to normalize each:

$$\Delta i_B(s) = \frac{\Delta i(s)}{-\sum_{k=1}^{B} P_k \log P_k} \quad . \tag{7}$$

And then we can again choose the feature that maximizes this normalized criterion.

When to Stop Splitting?

- If we continue to grow the tree until each leaf node has its lowest impurity (just one sample datum), then we will likely have over-trained the data. This tree will most definitely not generalize well.

When to Stop Splitting?

- If we continue to grow the tree until each leaf node has its lowest impurity (just one sample datum), then we will likely have over-trained the data. This tree will most definitely not generalize well.
- Conversely, if we stop growing the tree too early, the error on the training data will not be sufficiently low and performance will again suffer.

When to Stop Splitting?

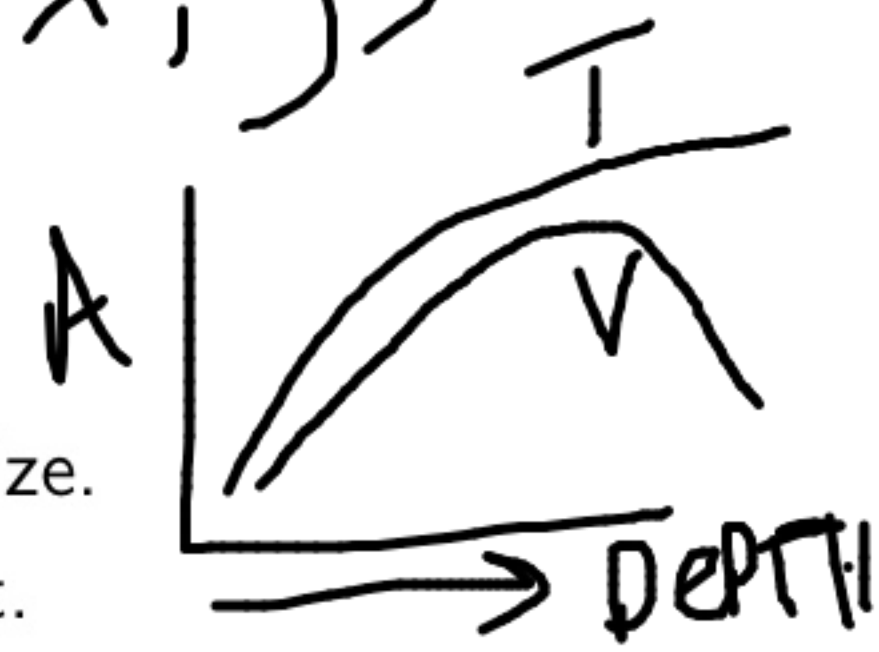
- If we continue to grow the tree until each leaf node has its lowest impurity (just one sample datum), then we will likely have over-trained the data. This tree will most definitely not generalize well.
- Conversely, if we stop growing the tree too early, the error on the training data will not be sufficiently low and performance will again suffer.
- So, how to stop splitting?

When to Stop Splitting?

- If we continue to grow the tree until each leaf node has its lowest impurity (just one sample datum), then we will likely have over-trained the data. This tree will most definitely not generalize well.
- Conversely, if we stop growing the tree too early, the error on the training data will not be sufficiently low and performance will again suffer.
- So, how to stop splitting?

1. Cross-validation...
2. Threshold on the impurity gradient.
3. Incorporate a tree-complexity term and minimize.
4. Statistical significance of the impurity gradient.

$D_{T_r} = \{x, y\}$



Stopping by Thresholding the Impurity Gradient

- Splitting is stopped if the best candidate split at a node reduces the impurity by less than the preset amount, β

$$\max_{s} \Delta i(s) \leq \beta \ . \tag{8}$$

Stopping by Thresholding the Impurity Gradient

- Splitting is stopped if the best candidate split at a node reduces the impurity by less than the preset amount, β:

$$\max_{s} \Delta i(s) \leq \beta \ . \tag{8}$$

- Benefit 1: Unlike cross-validation, the tree is trained on the complete training data set.

Stopping by Thresholding the Impurity Gradient

- Splitting is stopped if the best candidate split at a node reduces the impurity by less than the preset amount, β:

$$\max_s \Delta i(s) \leq \beta \ . \tag{8}$$

- Benefit 1: Unlike cross-validation, the tree is trained on the complete training data set.
- Benefit 2: Leaf nodes can lie in different levels of the tree, which is desirable whenver the complexity of the data varies throughout the range of values.

Stopping by Thresholding the Impurity Gradient

- Splitting is stopped if the best candidate split at a node reduces the impurity by less than the preset amount, β:

$$\max_s \Delta i(s) \leq \beta \quad . \tag{8}$$

- Benefit 1: Unlike cross-validation, the tree is trained on the complete training data set.
- Benefit 2: Leaf nodes can lie in different levels of the tree, which is desirable whenver the complexity of the data varies throughout the range of values.
- Drawback: But, how do we set the value of the threshold β?

Stopping with a Complexity Term

- Define a new global criterion function

$$\alpha \cdot \text{size} + \sum_{\text{leaf nodes}} i(N) \ . \tag{9}$$

which trades complexity for accuracy. Here, size could represent the number of nodes or links and α is some positive constant.

Stopping with a Complexity Term

- Define a new global criterion function

$$\alpha \cdot \text{size} + \sum_{\text{leaf nodes}} i(N) \quad . \tag{9}$$

  which trades complexity for accuracy. Here, size could represent the number of nodes or links and α is some positive constant.
- The strategy is then to split until a minimum of this global criterion function has been reached.

Stopping with a Complexity Term

- Define a new global criterion function

$$\alpha \cdot \text{size} + \sum_{\text{leaf nodes}} i(N) \ . \tag{9}$$

  which trades complexity for accuracy. Here, size could represent the number of nodes or links and α is some positive constant.
- The strategy is then to split until a minimum of this global criterion function has been reached.
- Given the entropy impurity, this global measure is related to the minimum description length principle.
  - The sum of the impurities at the leaf nodes is a measure of uncertainty in the training data given the model represented by the tree.

Stopping with a Complexity Term

- Define a new global criterion function

$$\alpha \cdot \text{size} + \sum_{\text{leaf nodes}} i(N) \ . \quad (9)$$

which trades complexity for accuracy. Here, size could represent the number of nodes or links and α is some positive constant.

- The strategy is then to split until a minimum of this global criterion function has been reached.
- Given the entropy impurity, this global measure is related to the minimum description length principle.
  - The sum of the impurities at the leaf nodes is a measure of uncertainty in the training data given the model represented by the tree.
- But, again, how do we set the constant α?

Stopping by Testing the Statistical Significance

- During construction, estimate the distribution of the impurity gradients Δi for the current collection of nodes.

Stopping by Testing the Statistical Significance

- During construction, estimate the distribution of the impurity gradients Δi for the current collection of nodes.
- For any candidate split, estimate if it is statistical different from zero. One possibility is the chi-squared test.

Stopping by Testing the Statistical Significance

- During construction, estimate the distribution of the impurity gradients Δi for the current collection of nodes.
- For any candidate split, estimate if it is statistical different from zero. One possibility is the chi-squared test.
- More generally, we can consider a hypothesis testing approach to stopping: we seek to determine whether a candidate split differs significantly from a random split.

Stopping by Testing the Statistical Significance

- During construction, estimate the distribution of the impurity gradients Δi for the current collection of nodes.
- For any candidate split, estimate if it is statistical different from zero. One possibility is the chi-squared test.
- More generally, we can consider a hypothesis testing approach to stopping: we seek to determine whether a candidate split differs significantly from a random split.
- Suppose we have n samples at node N. A particular split s sends Pn patterns to the left branch and $(1-P)n$ patterns to the right branch. A random split would place P_{n_1} of the ω_1 samples to the left, P_{n_2} of the ω_2 samples to the left and corresponding amounts to the right.

- The chi-squared statistic calculates the deviation of a particular split s from this random one:

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{iL} - n_{ie})^2}{n_{ie}} \tag{10}$$

where n_{iL} is the number of ω_1 patterns sent to the left under s, and $n_{ie} = Pn_i$ is the number expected by the random rule.

- The chi-squared statistic calculates the deviation of a particular split s from this random one:

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{iL} - n_{ie})^2}{n_{ie}} \tag{10}$$

where n_{iL} is the number of ω_1 patterns sent to the left under s, and $n_{ie} = Pn_i$ is the number expected by the random rule.

- The larger the chi-squared statistic, the more the candidate split deviates from a random one.

- The chi-squared statistic calculates the deviation of a particular split s from this random one:

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{iL} - n_{ie})^2}{n_{ie}} \tag{10}$$

  where n_{iL} is the number of ω_1 patterns sent to the left under s, and $n_{ie} = Pn_i$ is the number expected by the random rule.
- The larger the chi-squared statistic, the more the candidate split deviates from a random one.
- When it is greater than a critical value (based on desired significance bounds), we reject the null hypothesis (the random split) and proceed with s.

Pruning

- Tree construction based on "when to stop splitting" biases the learning algorithm toward trees in which the greatest impurity reduction occurs near the root. It makes no attempt to *look ahead* at what splits may occur in the leaf and beyond.

Pruning

- Tree construction based on "when to stop splitting" biases the learning algorithm toward trees in which the greatest impurity reduction occurs near the root. It makes no attempt to *look ahead* at what splits may occur in the leaf and beyond.
- **Pruning** is the principal alternative strategy for tree construction.

Pruning

- Tree construction based on "when to stop splitting" biases the learning algorithm toward trees in which the greatest impurity reduction occurs near the root. It makes no attempt to *look ahead* at what splits may occur in the leaf and beyond.
- **Pruning** is the principal alternative strategy for tree construction.
- In pruning, we exhaustively build the tree. Then, all pairs of neighboring leafs nodes are considered for elimination.

Pruning

- Tree construction based on "when to stop splitting" biases the learning algorithm toward trees in which the greatest impurity reduction occurs near the root. It makes no attempt to *look ahead* at what splits may occur in the leaf and beyond.
- **Pruning** is the principal alternative strategy for tree construction.
- In pruning, we exhaustively build the tree. Then, all pairs of neighboring leafs nodes are considered for elimination.
- Any pair that yields a satisfactory increase in impurity (a small one) is eliminated and the common ancestor node is declared a leaf.

Pruning

- Tree construction based on "when to stop splitting" biases the learning algorithm toward trees in which the greatest impurity reduction occurs near the root. It makes no attempt to *look ahead* at what splits may occur in the leaf and beyond.
- **Pruning** is the principal alternative strategy for tree construction.
- In pruning, we exhaustively build the tree. Then, all pairs of neighboring leafs nodes are considered for elimination.
- Any pair that yields a satisfactory increase in impurity (a small one) is eliminated and the common ancestor node is declared a leaf.
- Unbalanced trees often result from this style of pruning/merging.

Pruning

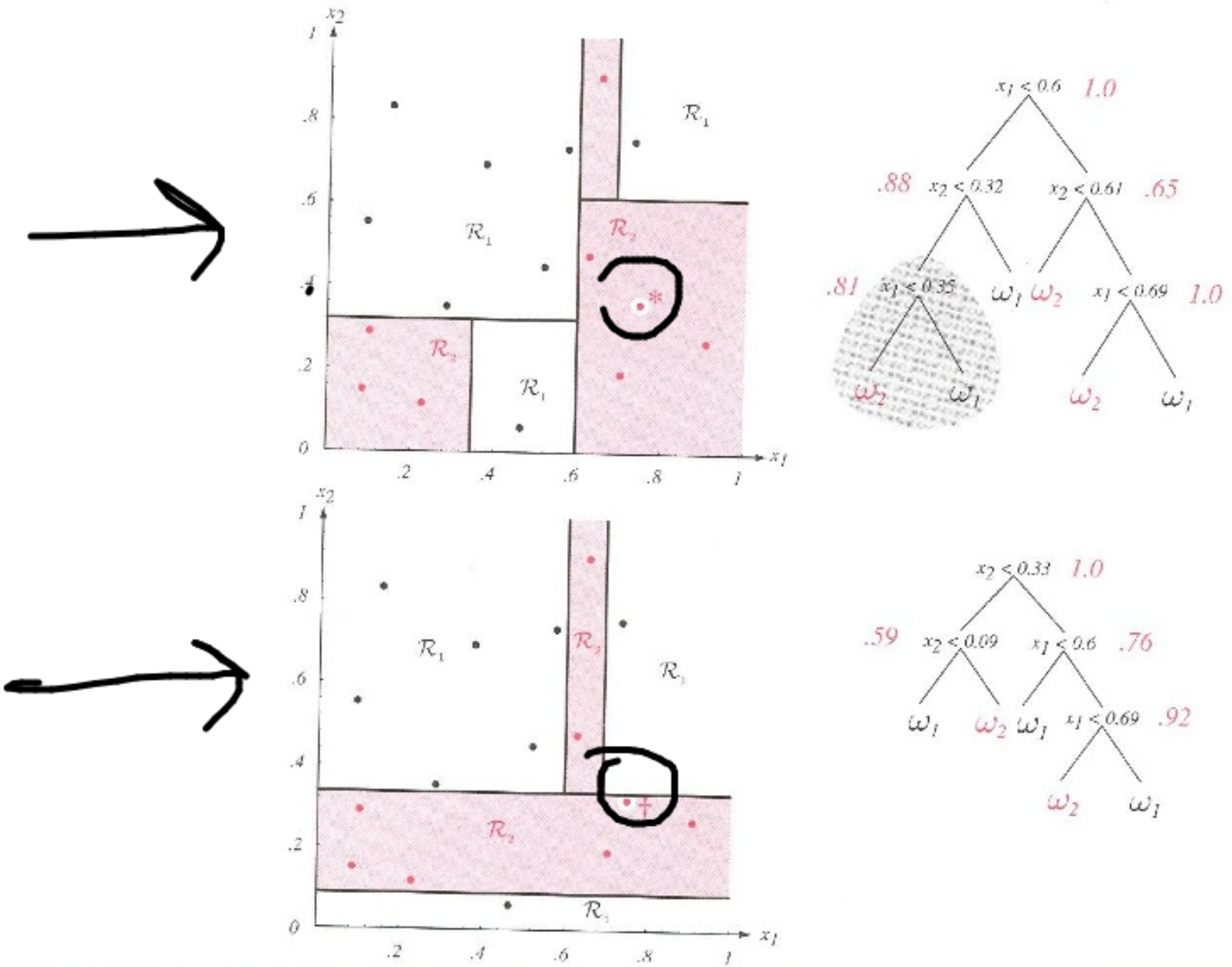
- Tree construction based on "when to stop splitting" biases the learning algorithm toward trees in which the greatest impurity reduction occurs near the root. It makes no attempt to *look ahead* at what splits may occur in the leaf and beyond.
- **Pruning** is the principal alternative strategy for tree construction.
- In pruning, we exhaustively build the tree. Then, all pairs of neighboring leafs nodes are considered for elimination.
- Any pair that yields a satisfactory increase in impurity (a small one) is eliminated and the common ancestor node is declared a leaf.
- Unbalanced trees often result from this style of pruning/merging.
- Pruning avoids the "local"-ness of the earlier methods and uses all of the training data, but it does so at added computational cost during the tree construction.

Assignment of Leaf Node Labels

- This part is easy...a particular leaf node should make the label assignment based on the distribution of samples in it during training. Take the label of the maximally represented class.
- We will see clear justification for this in the next chapter on Decision Theory.

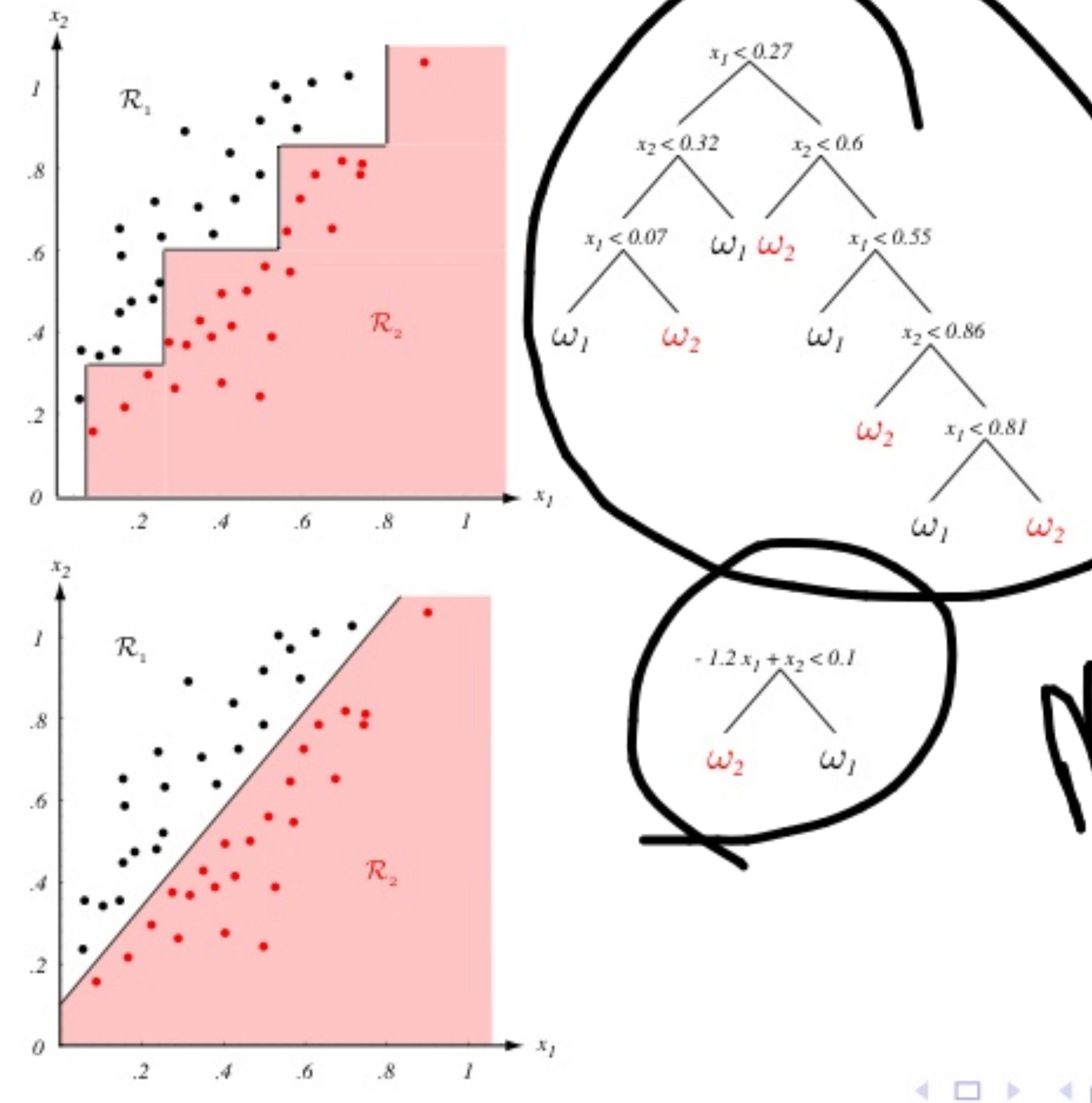
HARD

Instability of the Tree Construction



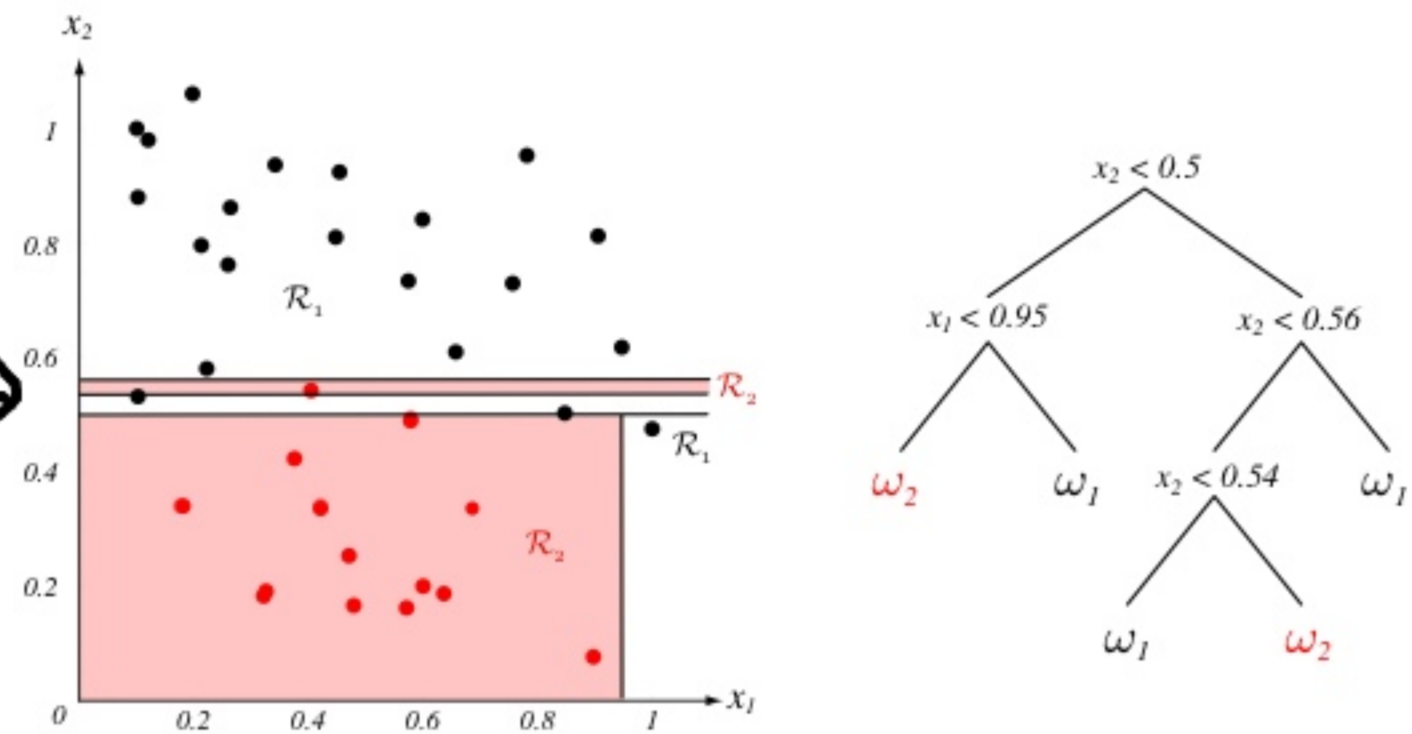
Importance of Feature Choice

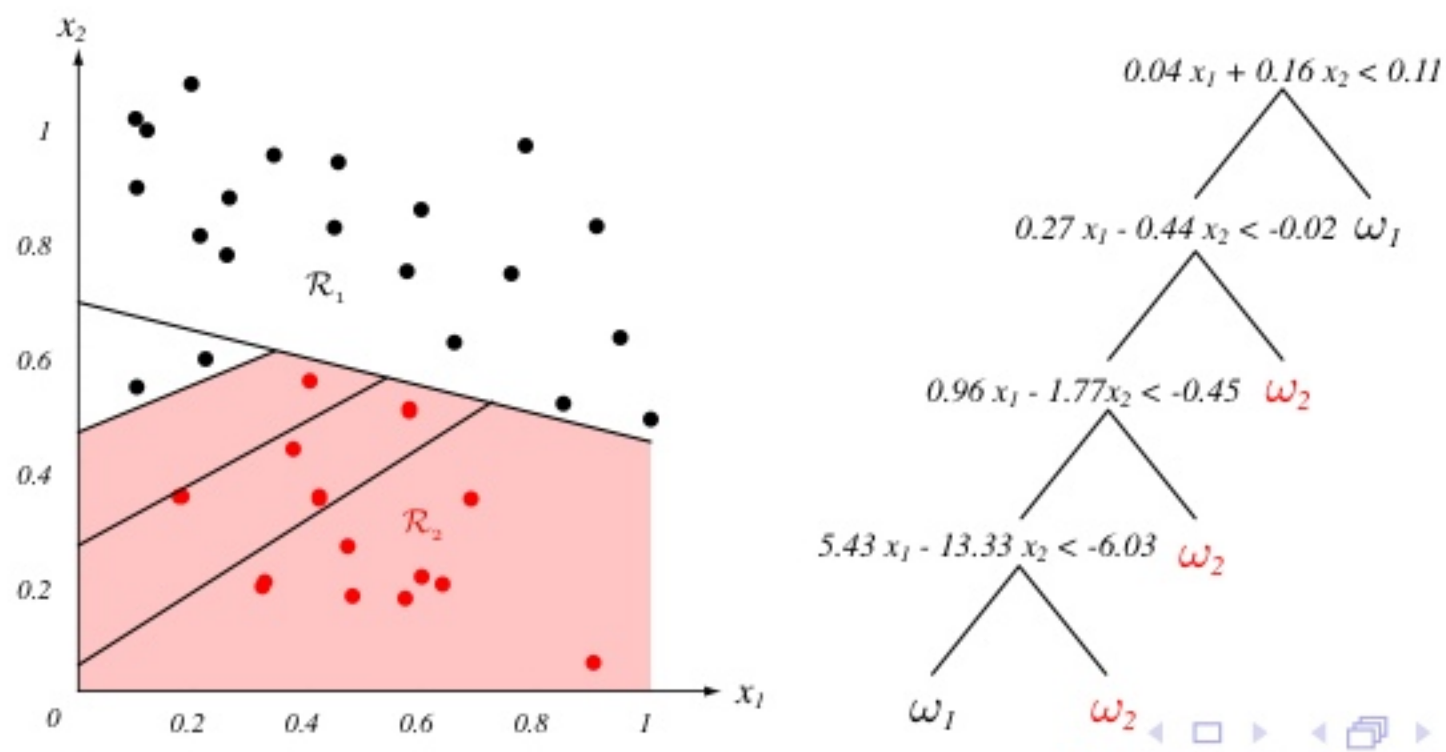
- The selection of features will ultimately play a major role in accuracy, generalization, and complexity.
- This is an instance of the Ugly Duckling principle.



model
Selection

- Furthermore, the use of multiple variables in selecting a decision rule may greatly improve the accuracy and generalization.





ID3 Method

- ID3 is another tree growing method.

ID3 Method

- ID3 is another tree growing method.
- It assumes nominal inputs.

ID3 Method

- ID3 is another tree growing method.
- It assumes nominal inputs.
- Every split has a branching factor B_j, where B_j is the number of discrete attribute bins of the variable j chosen for splitting.

ID3 Method

- ID3 is another tree growing method.
- It assumes nominal inputs.
- Every split has a branching factor B_j, where B_j is the number of discrete attribute bins of the variable j chosen for splitting.
- These are, hence, seldom binary.

ID3 Method

- ID3 is another tree growing method.
- It assumes nominal inputs.
- Every split has a branching factor B_j, where B_j is the number of discrete attribute bins of the variable j chosen for splitting.
- These are, hence, seldom binary.
- The number of levels in the trees are equal to the number of input variables.

ID3 Method

- ID3 is another tree growing method.
- It assumes nominal inputs.
- Every split has a branching factor B_j, where B_j is the number of discrete attribute bins of the variable j chosen for splitting.
- These are, hence, seldom binary.
- The number of levels in the trees are equal to the number of input variables.
- The algorithm continues until all nodes are pure or there are no more variables on which to split.

ID3 Method

- ID3 is another tree growing method.
- It assumes nominal inputs.
- Every split has a branching factor B_j, where B_j is the number of discrete attribute bins of the variable j chosen for splitting.
- These are, hence, seldom binary.
- The number of levels in the trees are equal to the number of input variables.
- The algorithm continues until all nodes are pure or there are no more variables on which to split.
- One can follow this by pruning.

C4.5 Method (in brief)

- This is a successor to the ID3 method.

C4.5 Method (in brief)

- This is a successor to the ID3 method.
- It handles real valued variables like CART and uses the ID3 multiway splits for nominal data.

C4.5 Method (in brief)

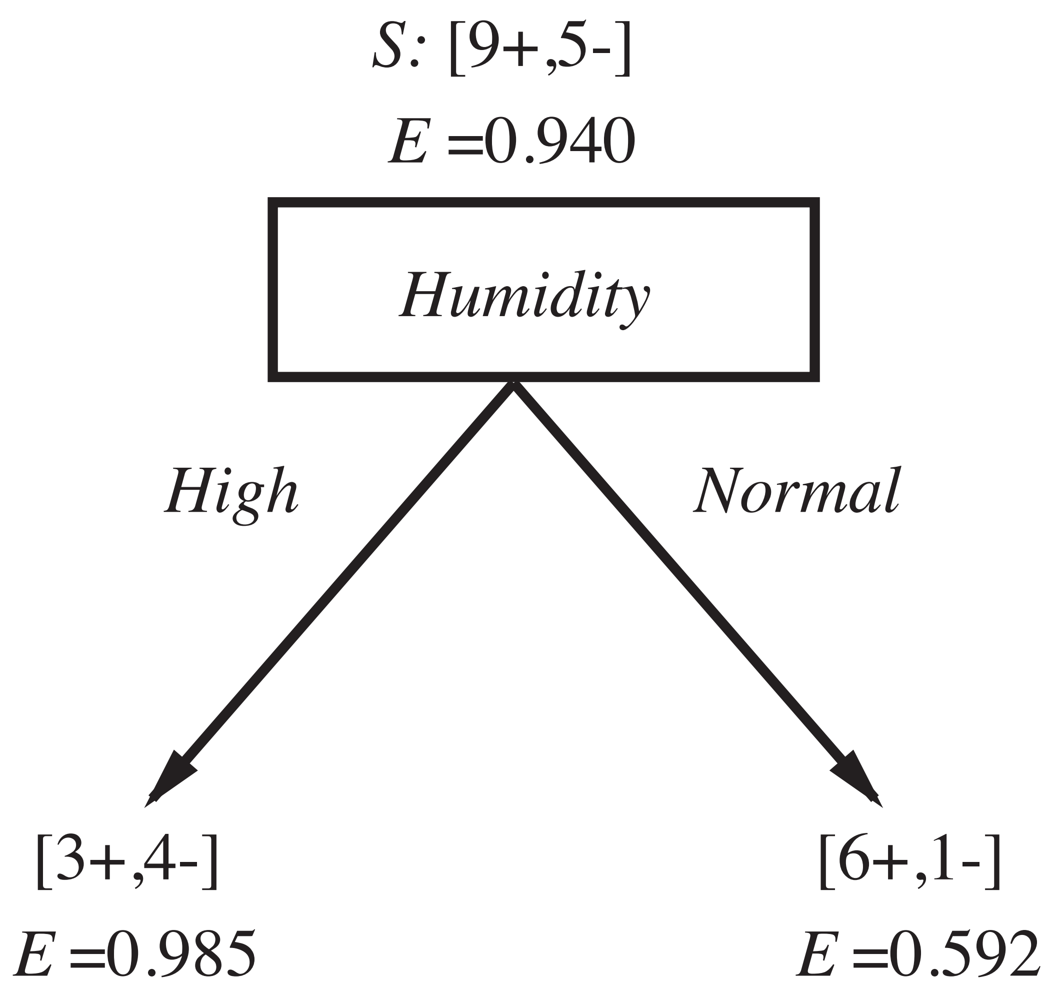
- This is a successor to the ID3 method.
- It handles real valued variables like CART and uses the ID3 multiway splits for nominal data.
- Pruning is performed based on statistical significance tests.

Quinlan

Example from T. Mitchell Book: PlayTennis

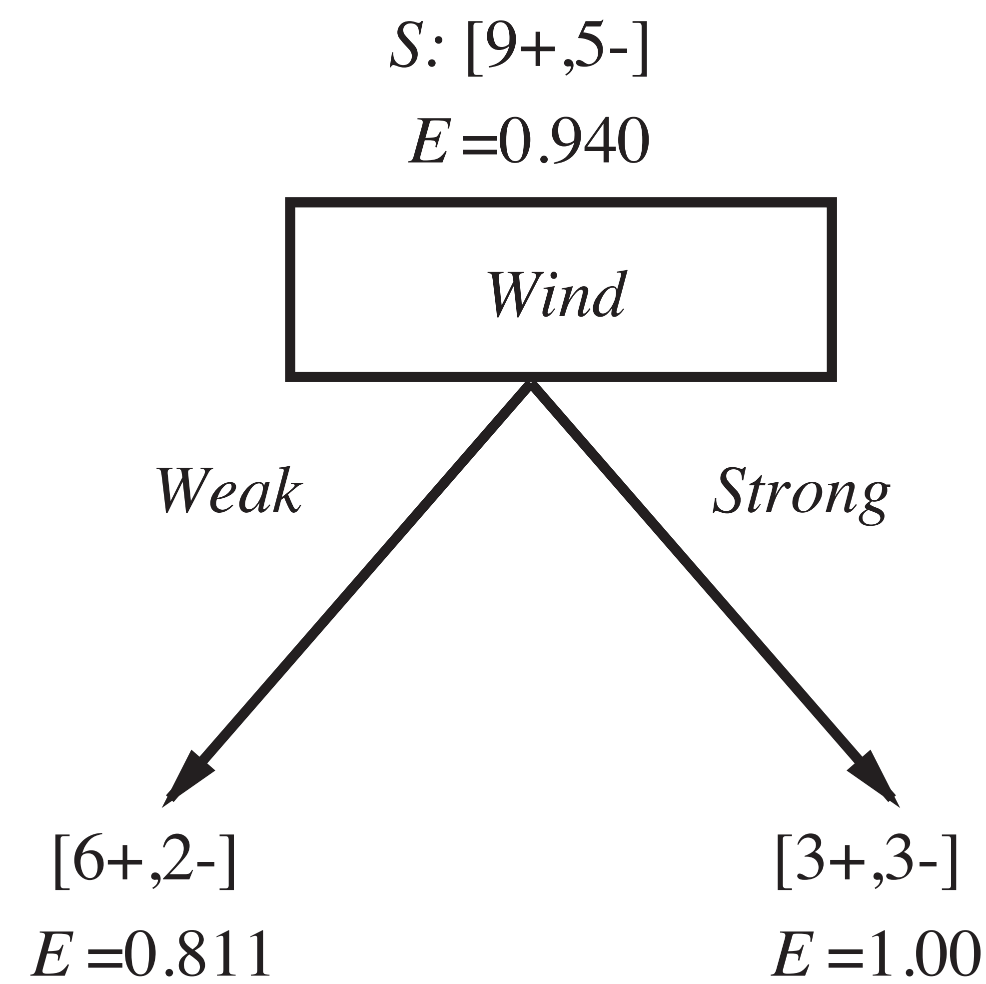
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	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 is the best classifier?



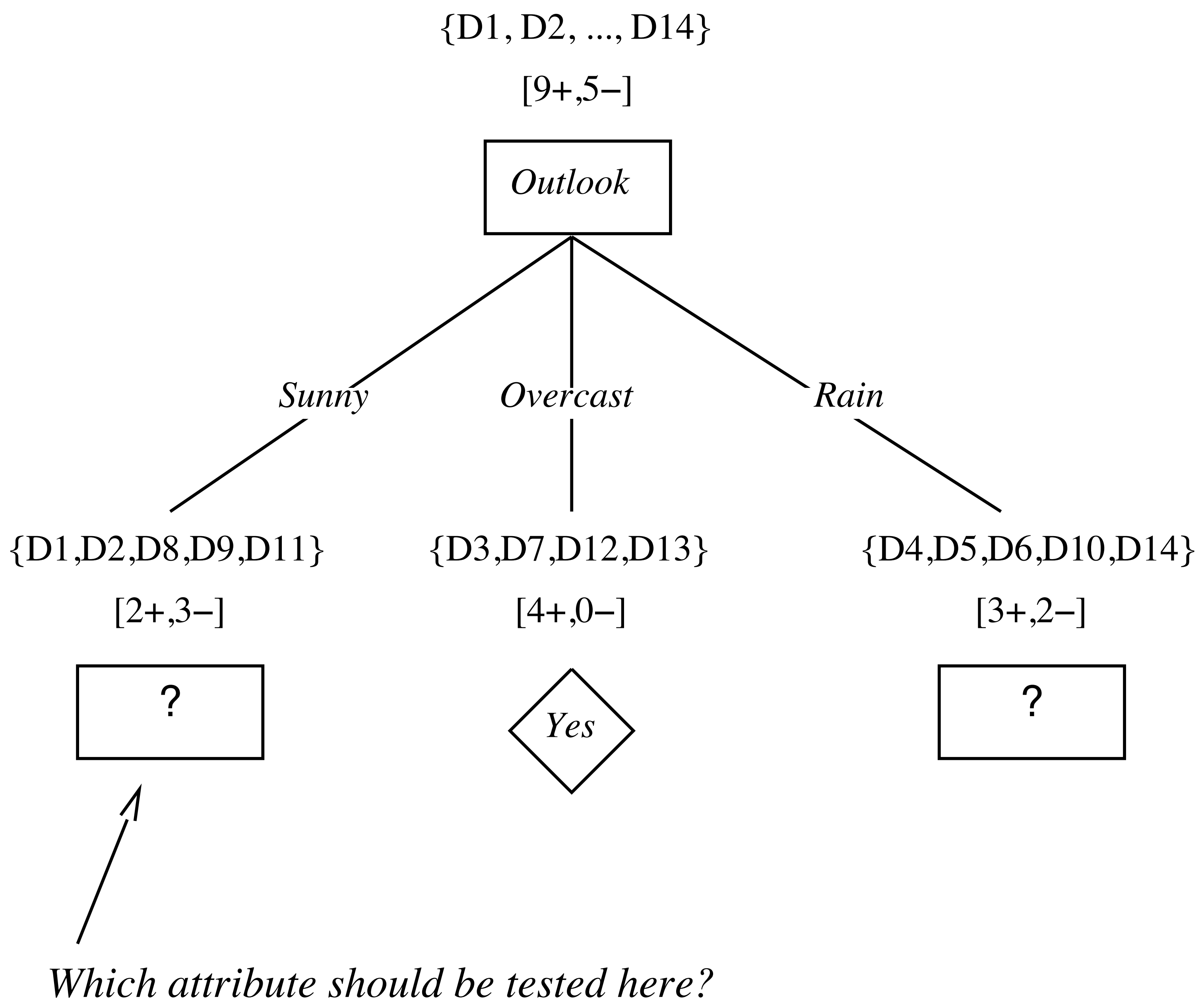
Gain (S, Humidity)

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



Gain (S, Wind)

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



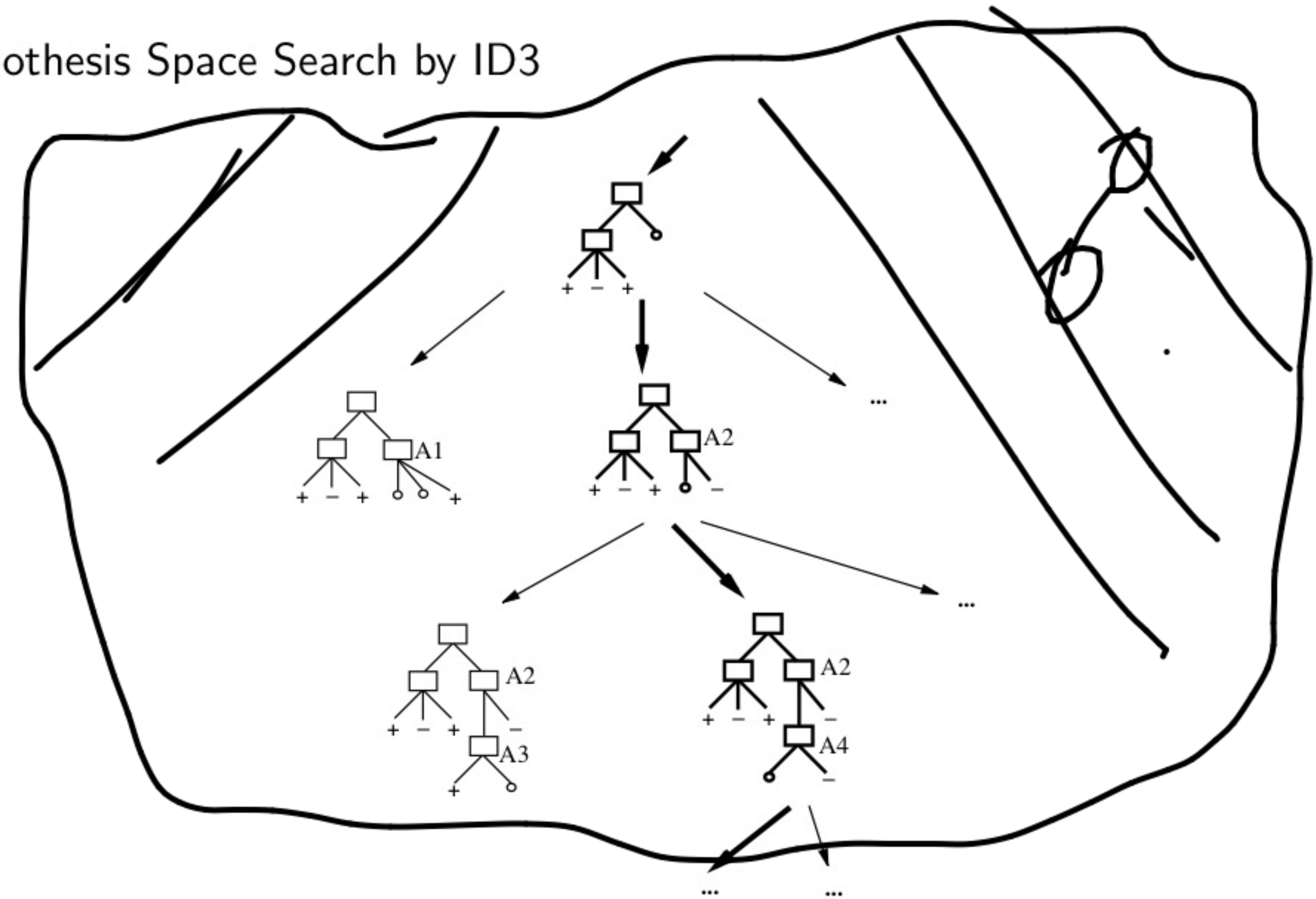
$S_{sunny} = \{D1, D2, D8, D9, D11\}$

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

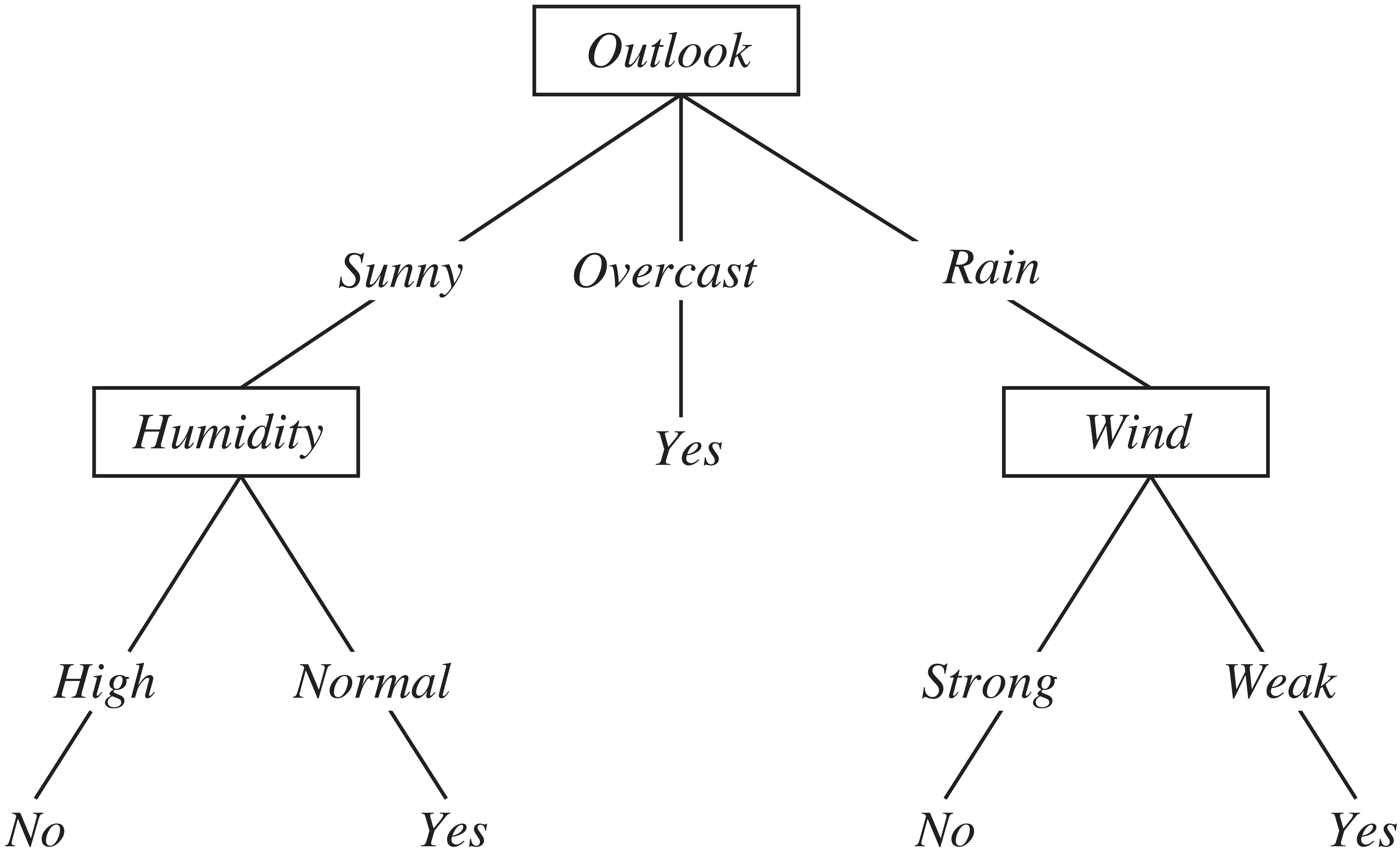
$$Gain\,(S_{sunny}, Temperature) = .970 - (2/5)\,0.0 - (2/5)\,1.0 - (1/5)\,0.0 = .570$$

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

Hypothesis Space Search by ID3



Learned Tree



Overfitting Instance

- Consider adding a new, noisy training example #15:

$$Sunny,\ Hot,\ Normal,\ Strong,\ PlayTennis = No$$

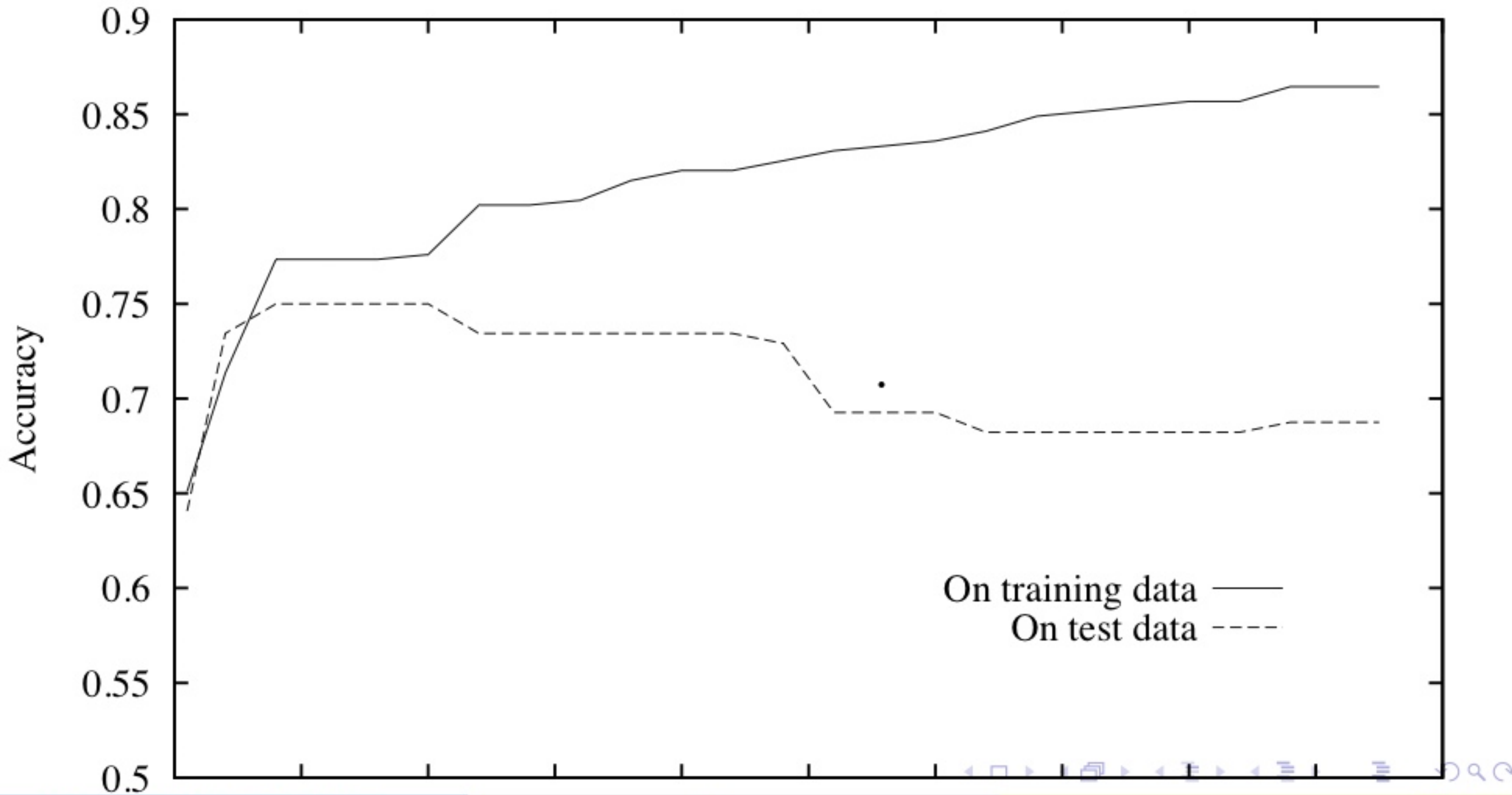
- What effect would it have on the earlier tree?

Overfitting Instance

- Consider adding a new, noisy training example #15:

$$Sunny,\ Hot,\ Normal,\ Strong,\ PlayTennis = No$$

- What effect would it have on the earlier tree?



Random Forest Classifiers Part I

1 Recap Decision Trees

Input $\mathcal{D} = \{(\mathbb{X}_1, y_1), (\mathbb{X}_2, y_2), \ldots, (\mathbb{X}_n, y_n)\}$

(note the categorical data) Assume $\mathbb{X}_i \in \{0,1\}^d$, $y_i \in \mathcal{L} = \{1, 2, \ldots, K\}$

$\mathbb{X} =$

note extension to $\mathbb{R}^d$

give example of $\{0,1\}^d$ Output: $f(\mathbb{X}) \xrightarrow{T} \mathcal{L}$ some classifier

Learn Decision Tree ($\mathcal{D}$, $T=0$, $p=$ ~~ϕ~~ root)

DFS

1. if $T >= $ MAXDEPTH or pure($\mathcal{D}$)
2. return ϕ
3. $q^* = \arg\max_{q \in \mathbb{X}} \Delta i(\mathcal{D}, q)$
4. $v = p.\mathrm{newChild}(q^*)$
5. $(L, R) = \mathrm{split}(\mathcal{D}, q^*)$
6. LearnDT(L, $T+1$, v)
7. LearnDT(R, $T+1$, v)

$\Delta i(\mathcal{D}, q) = i(\mathcal{D}) - P_L^{(q)} i(\mathcal{D}_L) - (1 - P_L(q))\, i(\mathcal{D}_R)$

$i(\mathcal{D}) = -\sum_j P(y_j) \log P(y_j)$

empirical probabilities from $\mathcal{D}$

2 Limitations of Decision Trees

(a) "Estimation Error" if d is big, need a lot of data

(b) What if d is huge or even dynamic?

Cannot evaluate Step 3

(c) Learning is greedy b/c of NP completeness of global learning

(d) Overfitting / reduce height

Random Forest Classifiers Part 1

3 Huge Features
Either dim is really big such as pairwise relation of pixels in an image or genes in bioinformatics
or for data $\mathbb{X}$ there is a function $G(\mathbb{X}) \rightarrow \{0,1\}^k$ generating some feature subset (not even countable)

Note "relation" to bagging

4 Randomized Tree Construction
→ idea is easy rather than consider all features when selecting a query at node v, consider only a small subset $\mathbb{X}_s$
* there are principled mathematics from Bayesian Decision Theory / information theory that underly these choices

⋮

3a $\mathbb{X}_s$ = Select Random Subset
3b $q^* = \arg\max_{q \in \mathbb{X}_s} \Delta i(\mathcal{D}, q)$
⋮

Proceed with the rest of the tree as before

Randomization overcomes feature problem and we can expect a large enough random tree to capture some discriminative information but insufficient for full classification

So

Random Forest Classifiers Part I

5 Random Forests

So, we can expect aggregating multiple tree responses into one classifier

Suppose we have a family of trees $T_1, \ldots, T_N$

$$f(\mathbf{x} \mid T_1, \ldots T_N) \to \mathcal{L}$$

Need a bit of probability here

draw examples on board

$P(y = c \mid T_i(v), \mathbf{x})$ posterior at that node v of T_i

Denote it μ_{T_v}

Aggregate $\bar{\mu}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \mu_{T_v}(\mathbf{x})$

For big $\mathcal{L}$, one may enforce sparsity in μ_{T_v}

$\hat{y} = \arg\max \bar{\mu}(\mathbf{x})$

max mode

555 2012
1/24

Random Tree Classifiers

Sources: ① Y Amit, D Geman "Shape Quantization and Recognition with Randomized Trees" Neural Computation 9,7 1545–1588, 1997

Limitations of Decision Trees

① Explore full feature space, but not practical when the space is too big — infinite dimensional

$\mathbb{X}$: space of digital images
$\mathcal{C}$: set of shape classes
Each image $\mathbf{x} \in \mathbb{X}$ has a true class label $Y(\mathbf{x}) \in \mathcal{C} = \{1, 2, \ldots, K\}$
Assume a probability dist. of X, $P(X)$
Goal is to construct a classifier $\hat{Y} : \mathbb{X} \to \mathcal{C}$ s.t. $P(\hat{Y} \neq Y)$ is small

---

Restrict to queries Q_A with bounded complexity, A with at most 20 tags and 20 relations. So $Q = \{Q_1, \ldots, Q_M\}$ be our feature vector
$Q \in \{0,1\}^M$
Training Set $\mathcal{L} = \{(\mathbf{x}_1, Y(\mathbf{x}_1)), \ldots, (\mathbf{x}_m, Y(\mathbf{x}_m))\}$