

# Recursive Partitioning

Lecture for CHM696D

10 March 2003

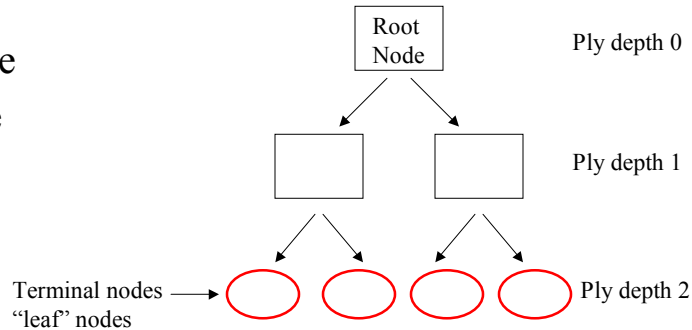
Rick Higgs & Dave Cummins  
Statistical & Information Sciences  
Lilly Research Laboratories

## What is Recursive Partitioning?

- RP is an exploratory technique for uncovering structure in data.
  - Models are constructed by successively splitting a dataset into increasingly homogeneous subsets until it is infeasible to continue, based on a set of "stopping rules."
  - The way the dataset is split is determined by the available covariates.
    - Notation: Covariate = Explanatory Variable = Independent Variable="Feature"
  - At each stage, all the covariates are examined and the one that gives the "best" split (e.g., greatest differences between groups) is chosen.
- What does an RP model look like?
  - RP generates a set of "rules." An example would be: "Potency is increased if there is a nitrogen attached to an aromatic ring."
  - Rules can be complex: "Potency is increased if there is a nitrogen attached to an aromatic ring *and* there is an oxygen present *and* at least two aromatic rings are present."

## Terminology of Tree Models is Graphic

- Nodes
- Root Node
- Terminal Nodes or Leaf Nodes
- Ply Depth
- Parent Node
- Child Node
- Path
- Pachinko



## Node Impurity Measures

- Continuous data: sample variance
- Categorical data: many possible choices
  - Binary data:
    - $\hat{p}(1 - \hat{p})$  binary variance
    - $\hat{p} \log(\hat{p}) + \hat{p} \log(1 - \hat{p})$  log likelihood
  - Multiple classes:
    - Pearson's chi-squared statistic
    - Maximum Entropy:  $-\sum_k \hat{p}(k) \log \hat{p}(k)$  (where  $0 \log 0 = 0$ )
    - Gini Index:  $\sum_{i \neq j} \hat{p}_i \hat{p}_j = 1 - \sum_j \hat{p}_j^2$
- From here forward consider the following:
  - Continuous response: t-test
  - Categorical response: chi-squared test

## Forward Stopping Rules

- P-value (or multiplicity-adjusted p-value) greater than a specified threshold (e.g. .05).
- Nodes become smaller than a specified minimum sample size per node (e.g. 5).
- Ply depth becomes greater than a specified value.

## Multiple Comparisons

- The RP algorithm as defined will be biased to select continuous valued features over binary features.
  - Multiple comparisons problem in statistics
- Bonferroni adjustment
  - First order correction for this problem
  - More sophisticated procedures are possible
  - Suppose we have 10 hypothesis tests
  - If we use  $\alpha = 0.05$  for each test then there is a 40% chance of spuriously rejecting hypothesis  $(1-(1-0.05)^{10}=.4)$
  - Want  $(1-\alpha')^k = (1-\alpha)$  where  $\alpha'$  is the level of each test
  - Bonferroni approximation: take  $\alpha' = \frac{\alpha}{k}$

## Pruning

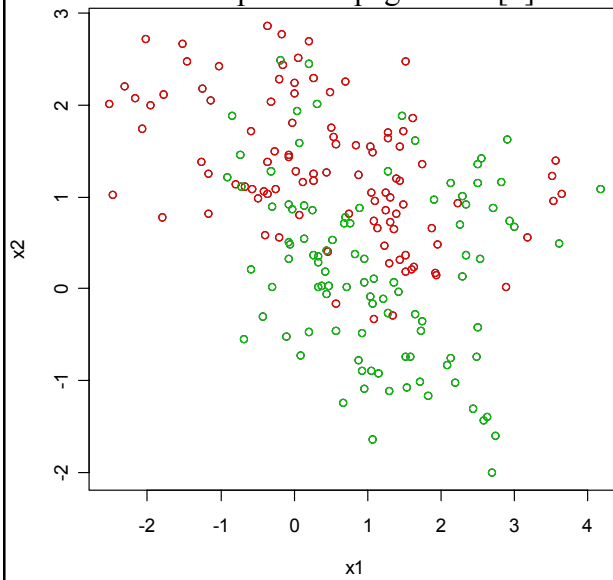
- Can be better to over-grow a tree and prune it back
- Forward stopping can miss good splits further down the tree
  - “horizon effect”
- Cost complexity pruning (Breiman) [4] popular
- Pruning less useful when data are more statistical in nature (e.g. noisy)
- Can be computationally burdensome
- Cross-validation required to set pruning parameter(s)

## Using the Tree to Predict

- "Pachinko" : new observations are subjected to the rules from the tree model based on their features. A new observation "falls" into a terminal node based on the rules.
- For regression: The mean response (from training data) for that node is taken to be the predicted value for the new observation. (This is why tree models never extrapolate.)
- For classification: The fractional class memberships in the node are taken as probability estimates of class memberships for the new observation.

# Two-Class Example

Example from page 17 of [1]

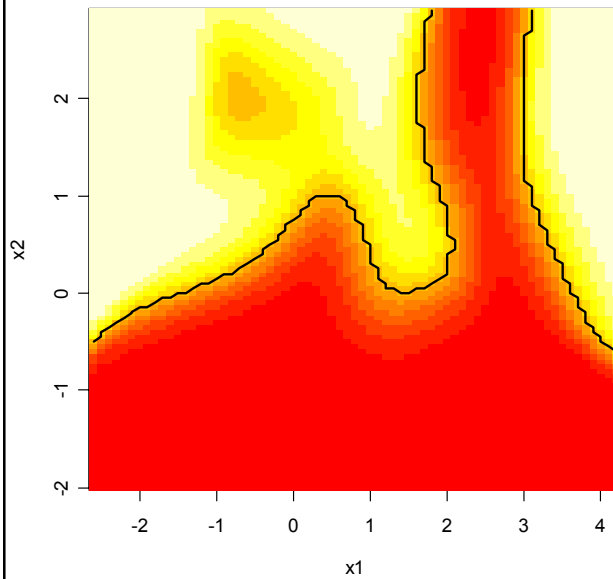


Class 0:  
10 Means from  $N([0,1]^2, I)$   
100 Samples  $\sim N(\mu_{0,i}, I/5)$

Class 1:  
10 Means from  $N([1,0]^2, I)$   
100 Samples  $\sim N(\mu_{1,i}, I/5)$

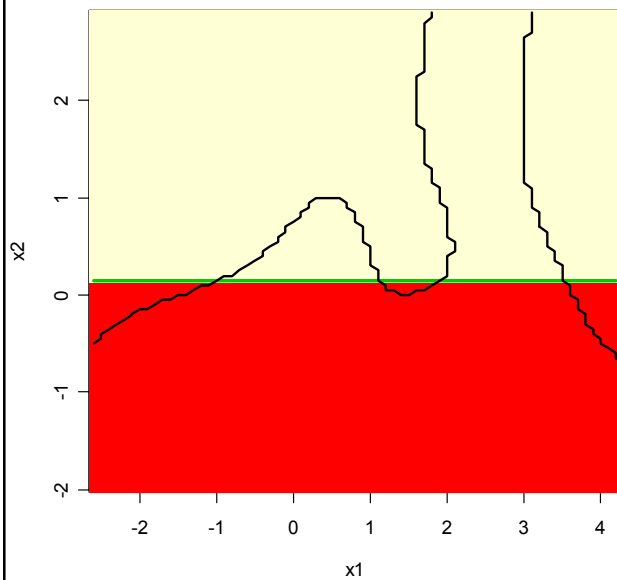
Test Set:  
Lattice  $x_1 = [-2.6 \text{ to } 4.2 \text{ by } 0.10]$   
 $x_2 = [-2.0 \text{ to } 2.9 \text{ by } 0.05]$   
6831 points total

# Two-Class Example



Color  $\sim P(\text{class 1} \mid \mathbf{x})$

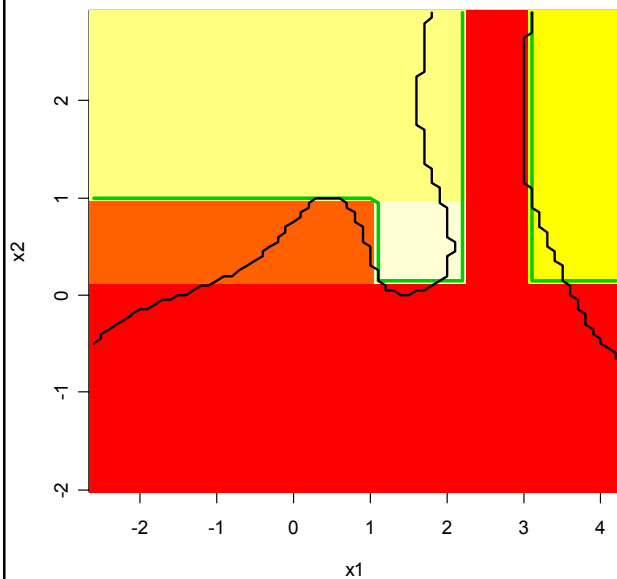
## Tree Model #1



Min Node Size = 10  
P-value = 0.001

Train Error = 28.5%  
Test Error = 29.4%  
Bayes Error = 19%

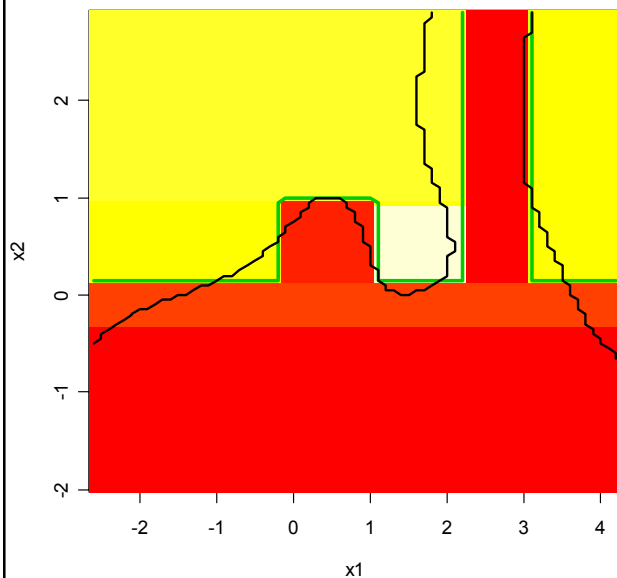
## Tree Model #2



Min Node Size = 5  
P-value = 0.05

Train Error = 14.5%  
Test Error = 25.0%  
Bayes Error = 19%

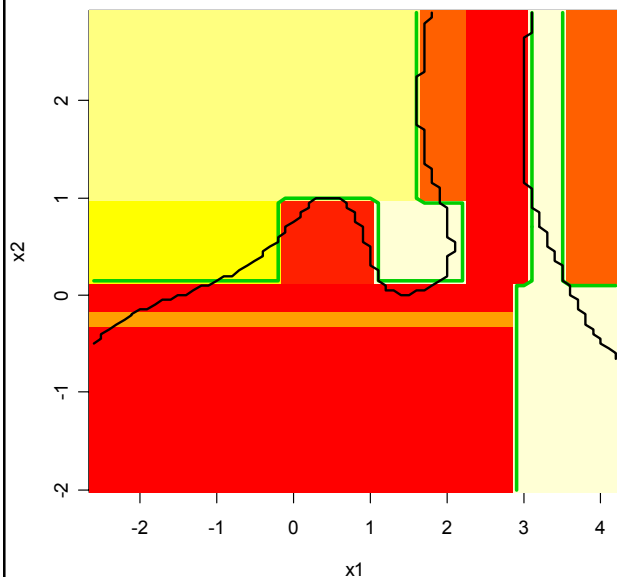
### Tree Model #3



Min Node Size = 3  
P-value = 0.5

Train Error = 13.5%  
Test Error = 23.1%  
Bayes Error = 19%

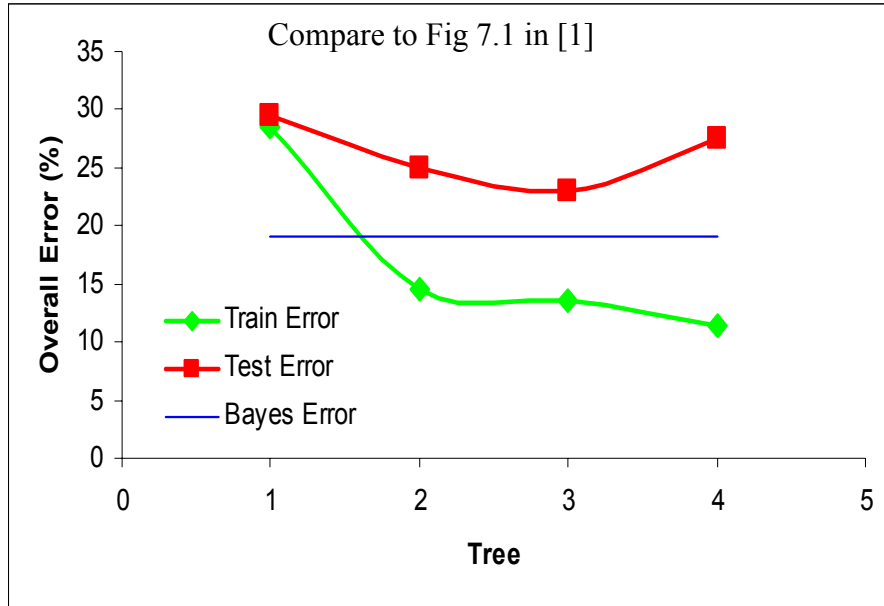
### Tree Model #4



Min Node Size = 1  
P-value = 1.0

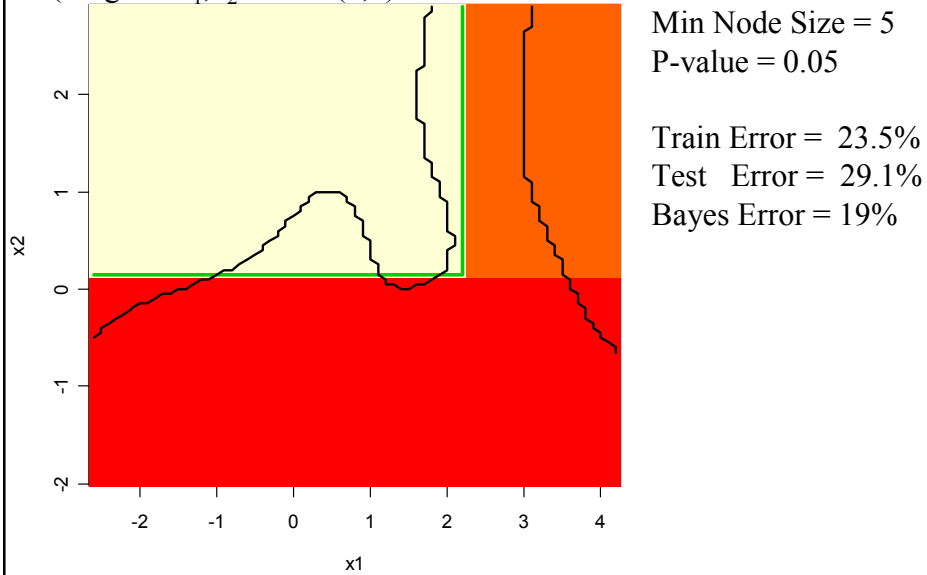
Train Error = 11.5%  
Test Error = 27.6%  
Bayes Error = 19%

## Tree Error Rates



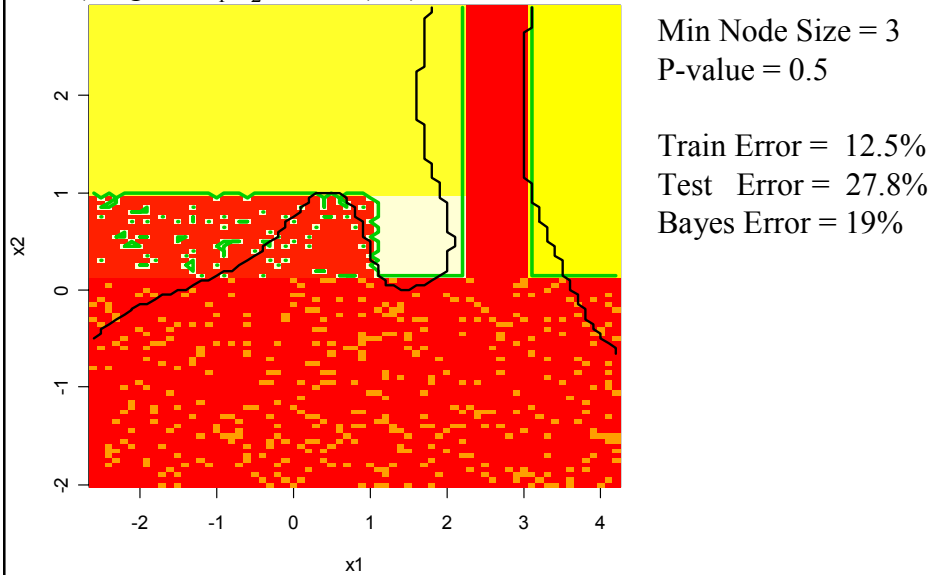
## Feature Selection

(Original  $x_1, x_2$  + 20  $U(0,1)$  Random Features)



## Feature Selection

(Original  $x_1, x_2$  + 20 U(0,1) Random Features)

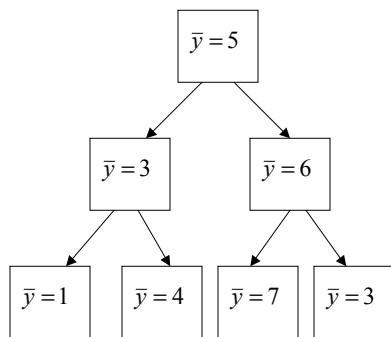


## Handling Missing Feature Values

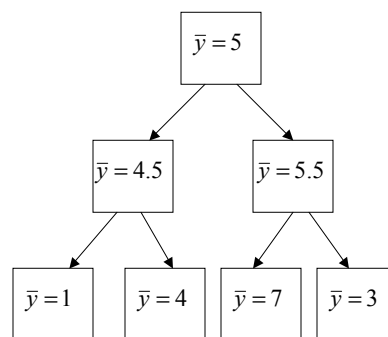
- Tree-based models deal with missing values in a local fashion. (locally, on-the-fly)
- Basic approaches
  - "Surrogate Splits" – exploits correlations in X
  - "Missing Value Buckets"
  - Drop the observation as far as it will go (naïve)
  - Remove observation all together (more naïve)
    - Don't do this, trees handle missing value nicely
  - Impute before building model

## Exploiting Interactions (automatically)

Easy to capture



Difficult to capture

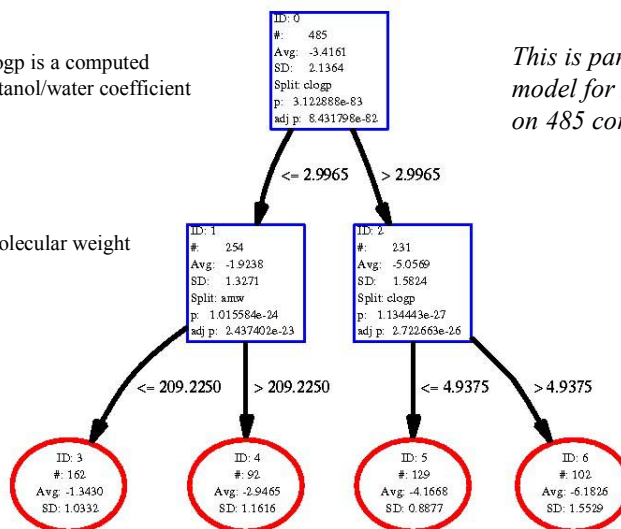


## Solubility Example

clogp is a computed octanol/water coefficient

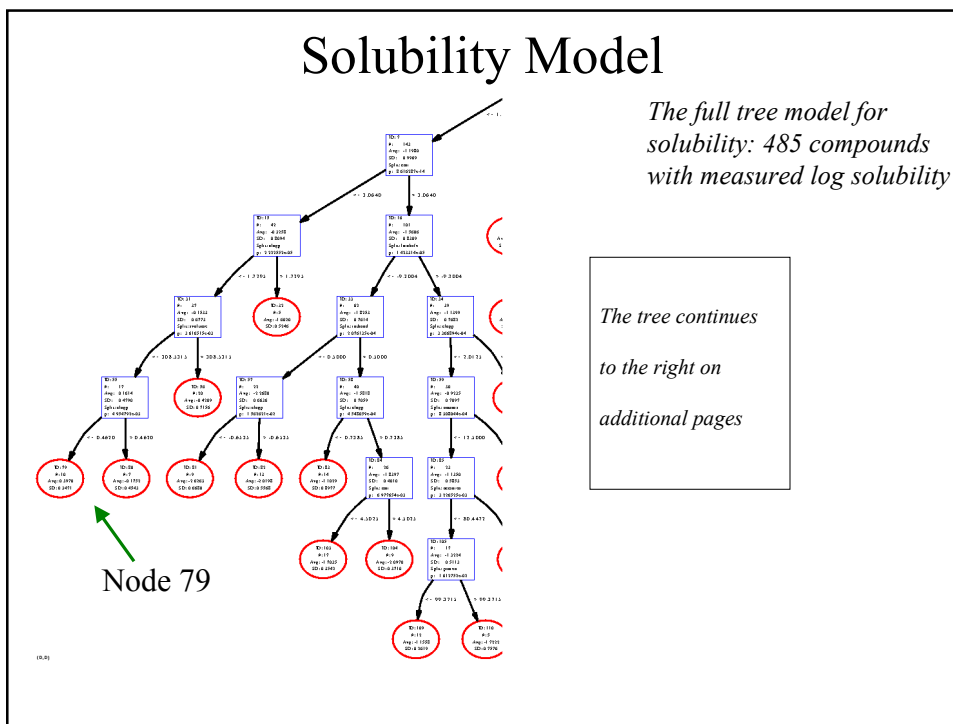
amw is molecular weight

*This is part of a tree model for solubility built on 485 compounds*



# Solubility Model

The full tree model for solubility: 485 compounds with measured log solubility

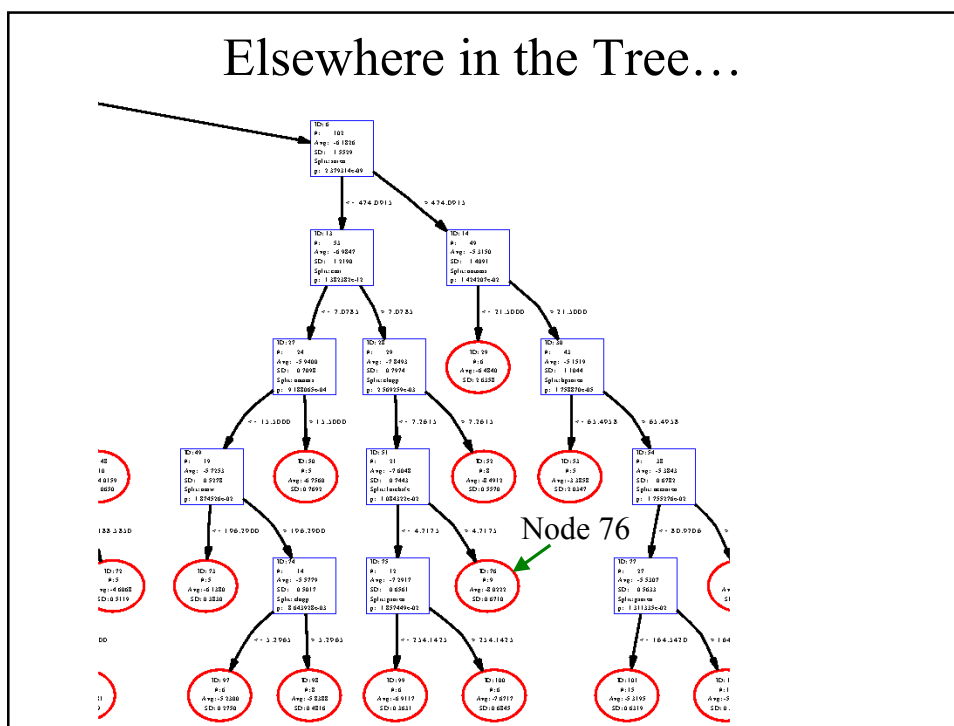


The tree continues to the right on additional pages

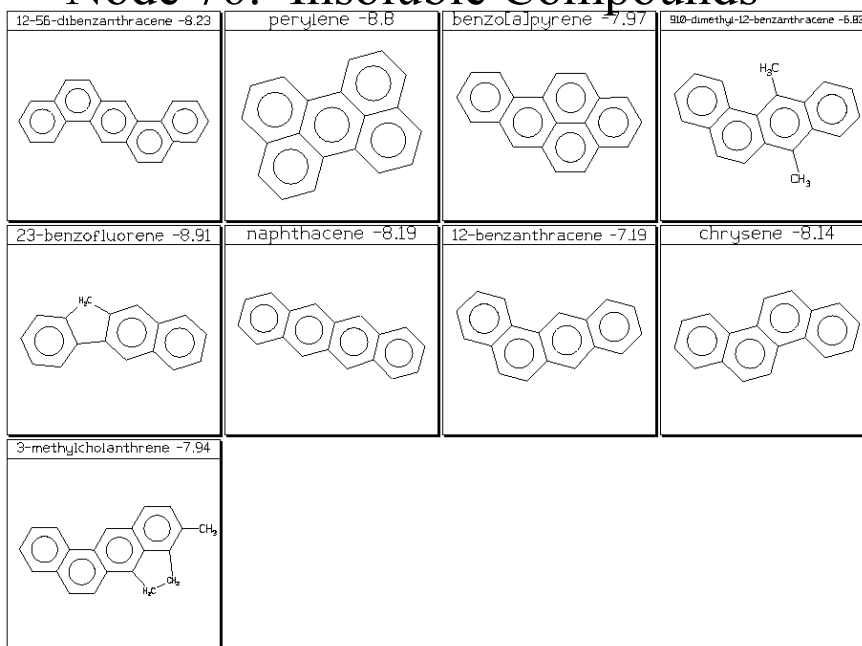
## Node 79: Soluble Compounds

oxalic_acid 0.389 <chem>OC(=O)C(=O)O</chem>	malonic_acid 0.762 <chem>OC(=O)CC(=O)O</chem>	amitrole 0.522 <chem>NC1=NC=NC=C1</chem>	acrylonitrile 0.145 <chem>C=CC#N</chem>
2-butanone 0.525 <chem>CC(=O)CC</chem>	nitromethane 0.255 <chem>C[N+](=O)[O-]</chem>	dicyanodiamide -0.311 <chem>NC(=N)NC#N</chem>	urea 0.946 <chem>NC(=O)N</chem>
glycine 0.493 <chem>NC(C(=O)O)</chem>			

# Elsewhere in the Tree...



## Node 76: Insoluble Compounds



## Node 79 Rules

- For the node 79, highly soluble cpds:
  - $\text{clogp} \leq 2.997$
  - $\text{amw} \leq 209$
  - free energy of solvation  $\leq 1.154$
  - $\text{cmr} \leq 3.064$
  - $\text{clogp} \leq 1.73$
  - $\text{svolume} \leq 308.53$
  - $\text{clogp} \leq 0.462$
- Note how  $\text{clogp}$  is re-used to refine the rule
- In this case the last  $\text{clogp}$  rule is the only  $\text{clogp}$  rule needed

## Node 76 Rules

- For the node 76, highly insoluble cpds:
  - $\text{clogp} > 2.997$
  - $\text{clogp} > 4.94$
  - $\text{sarea} \leq 474.09$
  - $\text{cmr} > 7.079$
  - $\text{clogp} \leq 7.262$
  - free energy of solvation  $> 4.717$
- Note how  $\text{clogp}$  is re-used to refine the rule
- In this case the consolidated  $\text{clogp}$  rule is:  
 $4.94 < \text{clogp} \leq 7.262$

## Error Rates

For a 2-class problem consider the 2x2 table:

		Truth	
		Class 0	Class 1
Predicted	Class 0	a	b
	Class 1	c	d

“Sensitivity” =  $P(\text{predict 1} \mid \text{truly 1}) = d / (b+d)$

“Specificity” =  $P(\text{predict 0} \mid \text{truly 0}) = a / (a+c)$

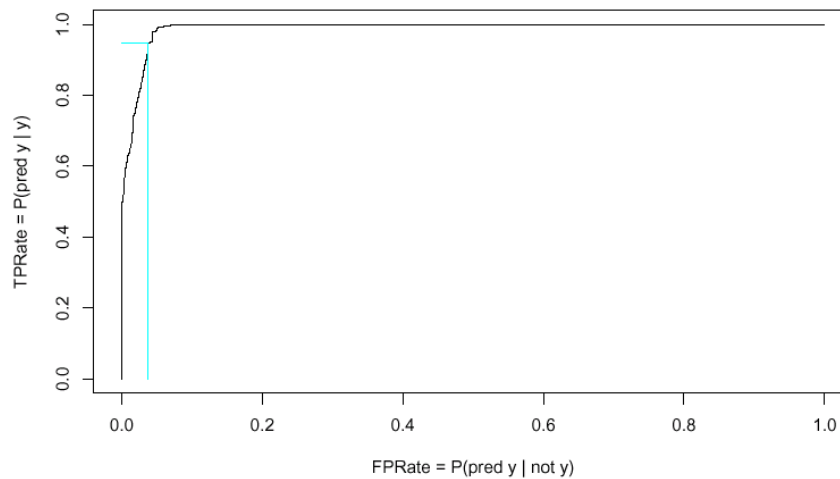
“False Positive Rate” =  $P(\text{predict 1} \mid \text{truly 0}) = 1 - \text{“Specificity”}$

“False Negative Rate” =  $P(\text{predict 0} \mid \text{truly 1}) = 1 - \text{“Sensitivity”}$

## ROC Curves

Example from a QSAR application

ROC AUC = 0.9893  
Optimal Score Threshold = 0.227



## Error Rates

- For some applications we are primarily interested in  $P(\text{class 1} \mid \text{predict class 1})$ .
- For example, consider the QSAR application where we build a predictive model for enzyme inhibition. We then use the model to score a large database of compounds we could purchase and test in the assay. Should we buy any of the compounds?

$P(\text{compound is active} \mid \text{compound predicted active})$

## Error Rate Example

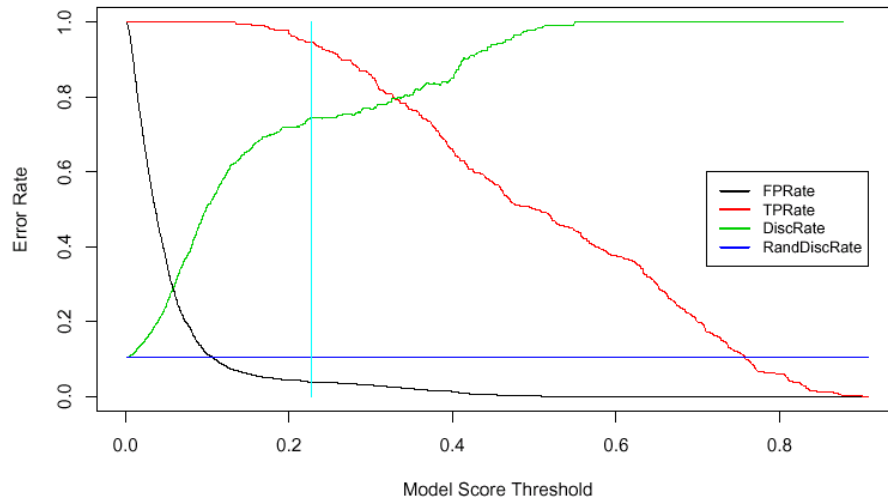
Prostate cancer PSA test: Sensitivity = 99.4% Specificity = 83.4%  
You're an otherwise healthy male and your PSA test comes back +.  
The prevalence of prostate cancer is 0.001 in your age group.  
What's the probability you have prostate cancer?

$$\begin{aligned} P(\text{cancer} \mid \text{predict cancer}) &= \frac{P(\text{predict cancer} \mid \text{cancer})P(\text{cancer})}{P(\text{predict cancer})} \\ &= \frac{\text{Sensitivity } P(\text{cancer})}{\text{Sensitivity } P(\text{cancer}) + (1 - \text{Specificity}) P(\text{no cancer})} \\ &= 0.006 \text{ (or 0.6\%)} \end{aligned}$$

# Error Rates

Example from a QSAR application

**Error Rates**  
At Optimal Thresh TP=0.948 FP=0.038 DR=0.745



## Recursive Partitioning Strengths

### **Recursive Partitioning:**

Does automatic variable selection.

Does not suffer from ill-conditioning problems. (no matrix inversion)

Captures nonlinear relationships automatically.

Gives intuitive and interpretable models in the form of "rules".

Captures and corrects for local effects (interactions).

Is robust to sparse missing values.

Is non-parametric.

Invariant to monotone transformations of explanatory variables.

Has the potential to automatically handle the mixture problem.

Scales linearly with the number of covariates.

Works in modular fashion; good for distributed implementation.

Conceptually simple, yet powerful

## Recursive Partitioning Weaknesses

### **Recursive Partitioning:**

Is unable to extrapolate beyond the range of observed responses.

Is less efficient than parametric models at revealing smooth trends.

Gives a non-parsimonious description of additive models.

Tends to require more observations than linear models.

- While reasonable RP models have been built on as few as 50 observations, an "interesting" model usually requires hundreds.

Interpretation can be misleading.

- A tree's simplistic form may fool a user into missing highly correlated but more relevant covariates.
- Different trees can often describe the same data.

Higher order interactions can be masked by nonsignificant main effects where two or more covariates perfectly counterbalance each other and so an initial split is not made.

Discrete, non-smooth nature of model (bigger problem for regression)

Unstable → high variance (slight change in data produces very different tree)

## Selected References

- (1) Hastie, T., Tibshirani, R. Friedman, J. (2001). *The Elements of Statistical Learning Theory*, Springer-Verlag, New York.
- (2) Duda, R., Hart, P. and Stork, D. (2000). *Pattern Classification (Second Edition)*, Wiley, New York.
- (3) Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.
- (4) Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth.