Data Mining

Tree Model

최 대 우
한국외국어대학교 통계학과

Agenda

- Introduction
- Growing Trees
- Pruning in CART
- Surrogate
- Tree regression
- Running S-PLUS

Example

- The figure 1 and 2 is the result of tree model in S-PLUS.
- The kyphosis data has 81 rows representing data on 81 children who have had corrective spinal surgery.
- The outcome Kyphosis is a binary variable, the other three variables (columns) are numeric.

Example (cont.)

- DATA DESCRIPTION
  - Kyphosis: a factor telling whether a postoperative deformity (kyphosis) is “present" or “absent" .
  - Age: the age of the child in months.
  - Number: the number of vertebrae involved in the operation.
  - Start: the beginning of the range of vertebrae involved in the operation.

Tree model with `kyphosis` (1)

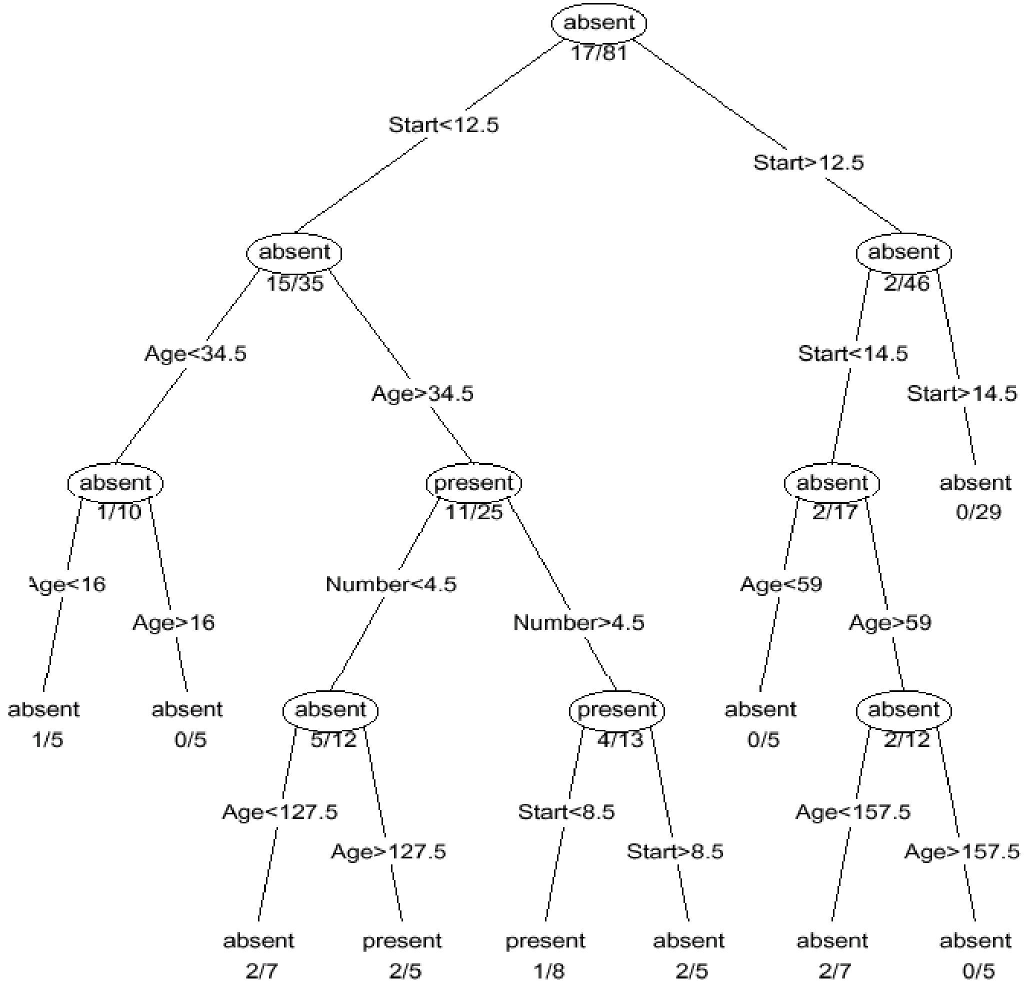


Figure 1 : Splitted by deviance

Tree model with `kyphosis` (2)

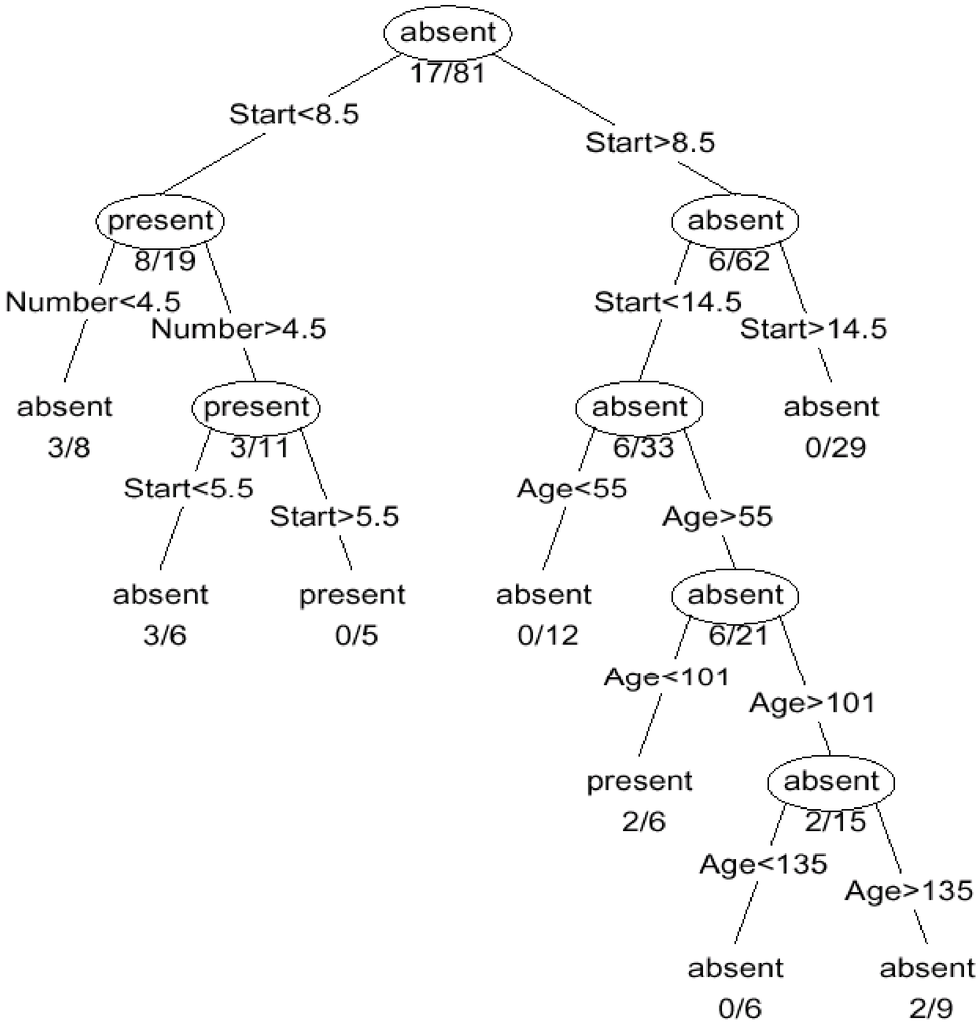
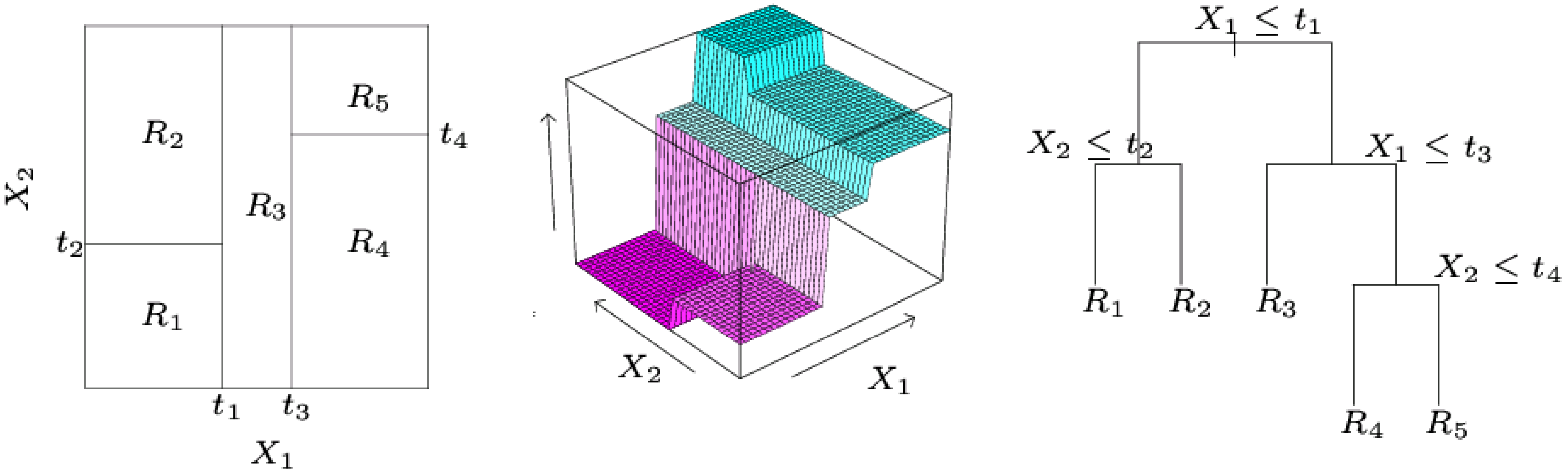
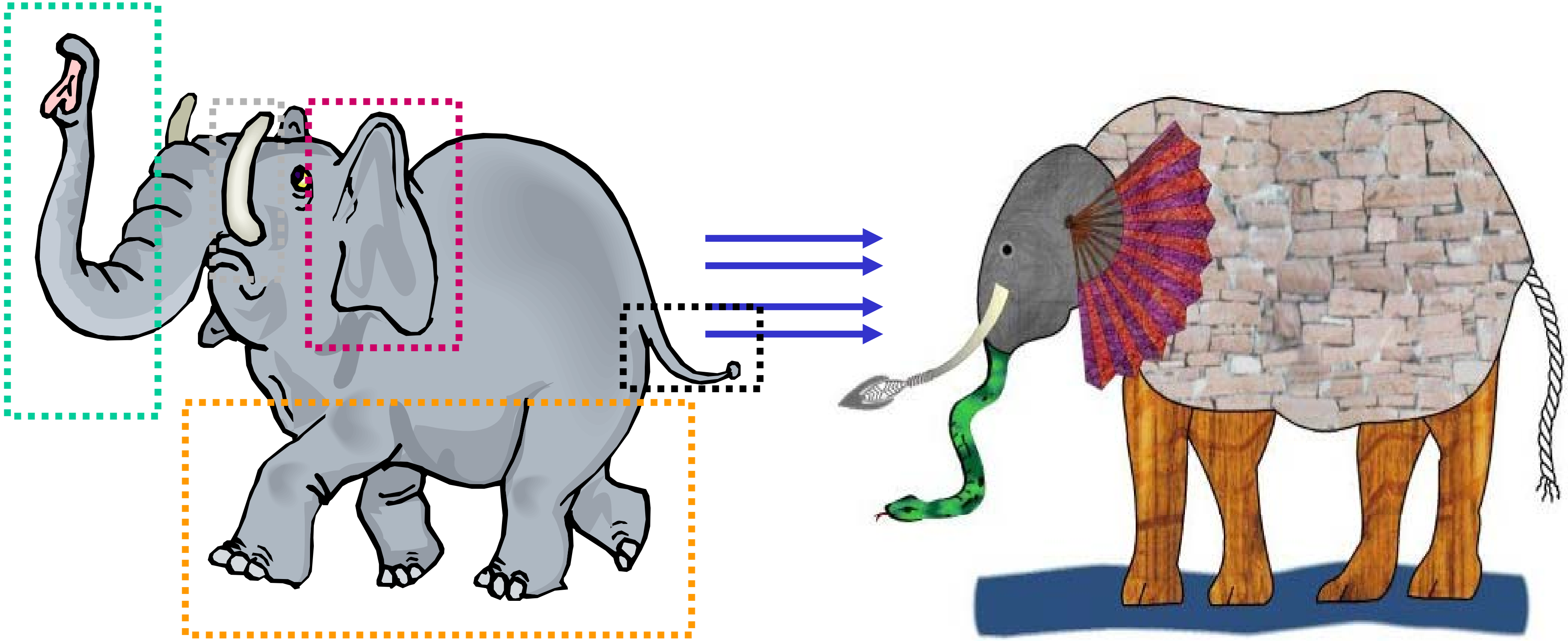


Figure 2: Splitted by Gini index

Tree Regression in CART



How CART Sees An Elephant



It was six men of Indostan ; To learning much inclined, Who went to see the Elephant; (Though all of them were blind), That each by observation; Might satisfy his mind ….

-- "**The Blind Men and the Elephant**" by John Godfrey Saxe (1816-1887)

Several approaches

- THAID Morgan and Messenger, 1973)
- CHAID (Kass, 1980)
- ID3 (Quinlan, 1983)
- CART (Breiman, Freidman, Olshen and Stone, 1984)
- FACT (Loh and Vanichetakul, 1988)
- CN2 (Clark and Niblett, 1989)
- C4.5 (Quinlan, 1993)
- QUETS (Loh and Shih, 1997)
- Ltree, Btree, Oblique Tree (Gama, 1997)

Some history

- Artificial intelligence – rules for exact prediction
  - Hunt, Marin, and Stone (1996)
  - Quinlan : ID3, C4.5 (1982, 1993) (classification only)
- Statistics – uncertainty
  - Sonquist and Morgan (1964)
  - Breiman, Freidman, Olshen and Stone : CART (1984)

Pros and Cons

- Pros - Why on earth would you want to fit such a model?
  - Easy to understand and interpret - non-statisticians really like it
  - Handles missing values efficiently
  - Fast computation
  - Interactions
  - invariant to monotone transforms

Pros and Cons (cont.)

- Cons
  - emphasizes interactions
  - non-parsimonious description of additive models
  - Prediction surface is not smooth.
  - Different trees can often describe the same data.
  - very unstable (high variance)
  - Accuracy is lower than other methods.

Growing Trees (1)

- Trees are grown recursively. A terminal node g is split into the left and right daughters (say, g_L and g_R) that increase the split criterion

$$D_g - D_{gL} - D_{gR}$$

highly, where D is a measure of goodness of fit

- Choice of goodness-of-fit measure:
    - for regression:

$$D_g = \sum_{i \in g} (y_i - \bar{y}_g)^2$$

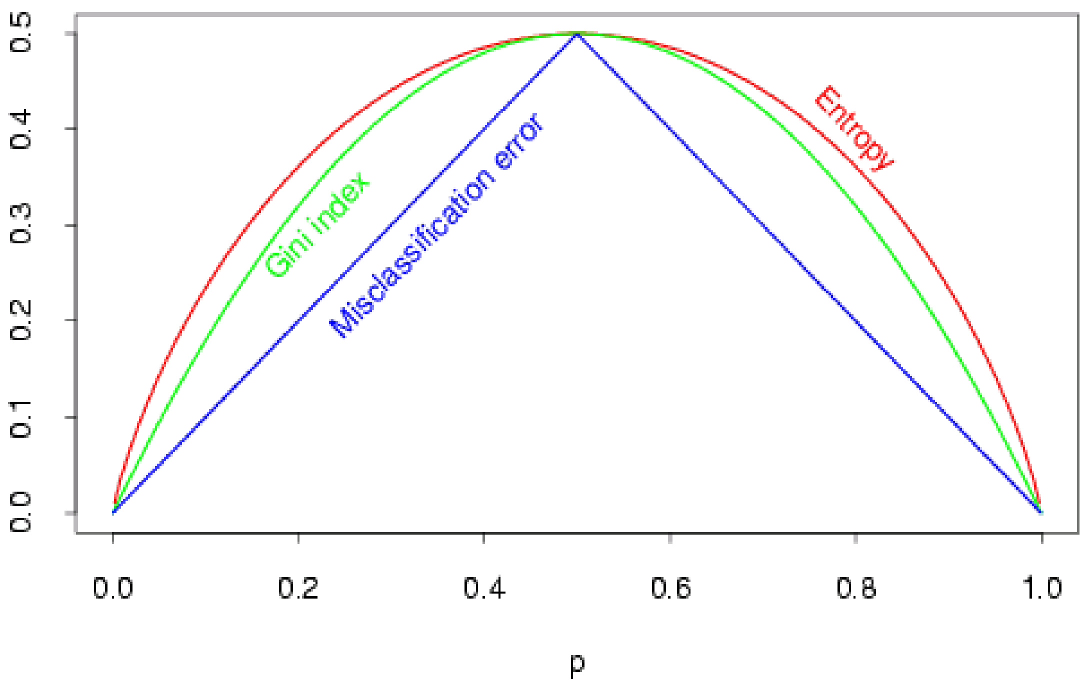
    - for classification:

Growing Trees (2)

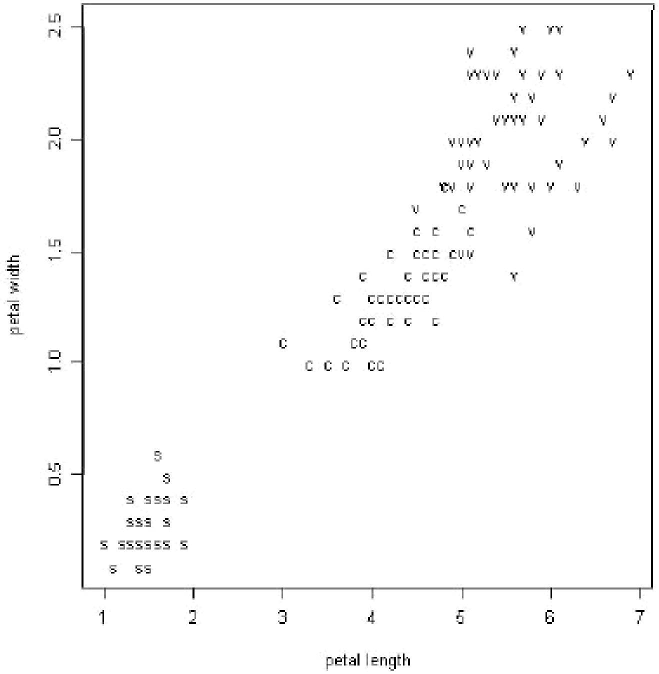
- Gini index :
  - $$D_g = \sum_j \hat{p}_j (1 - \hat{p}_j)$$ (CART)
- Entropy index :
  - $$D_g = -\sum_j \hat{p}_j \log \hat{p}_j$$ (C 4.5)
- Deviance :
  - $$D_g = -2\sum_j n_j \log \hat{p}_j$$ (S-PLUS)

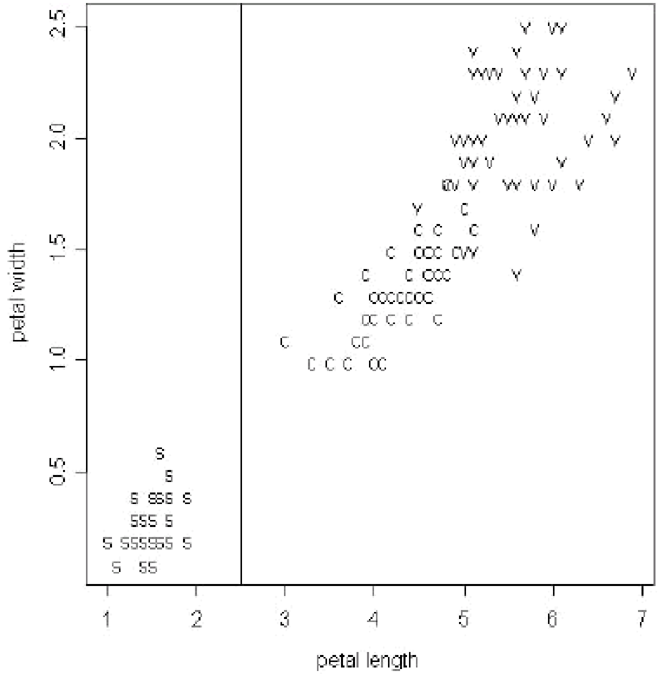
Criteria for splitting

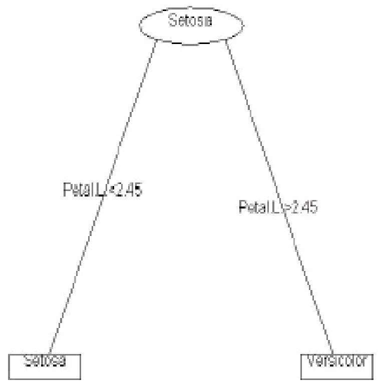
- Cross-entropy and Gini are more sensitive
- To grow the tree: use CE or Gini
- To prune the tree: use Misclassification rate (or any other method)



Growing Trees (4)

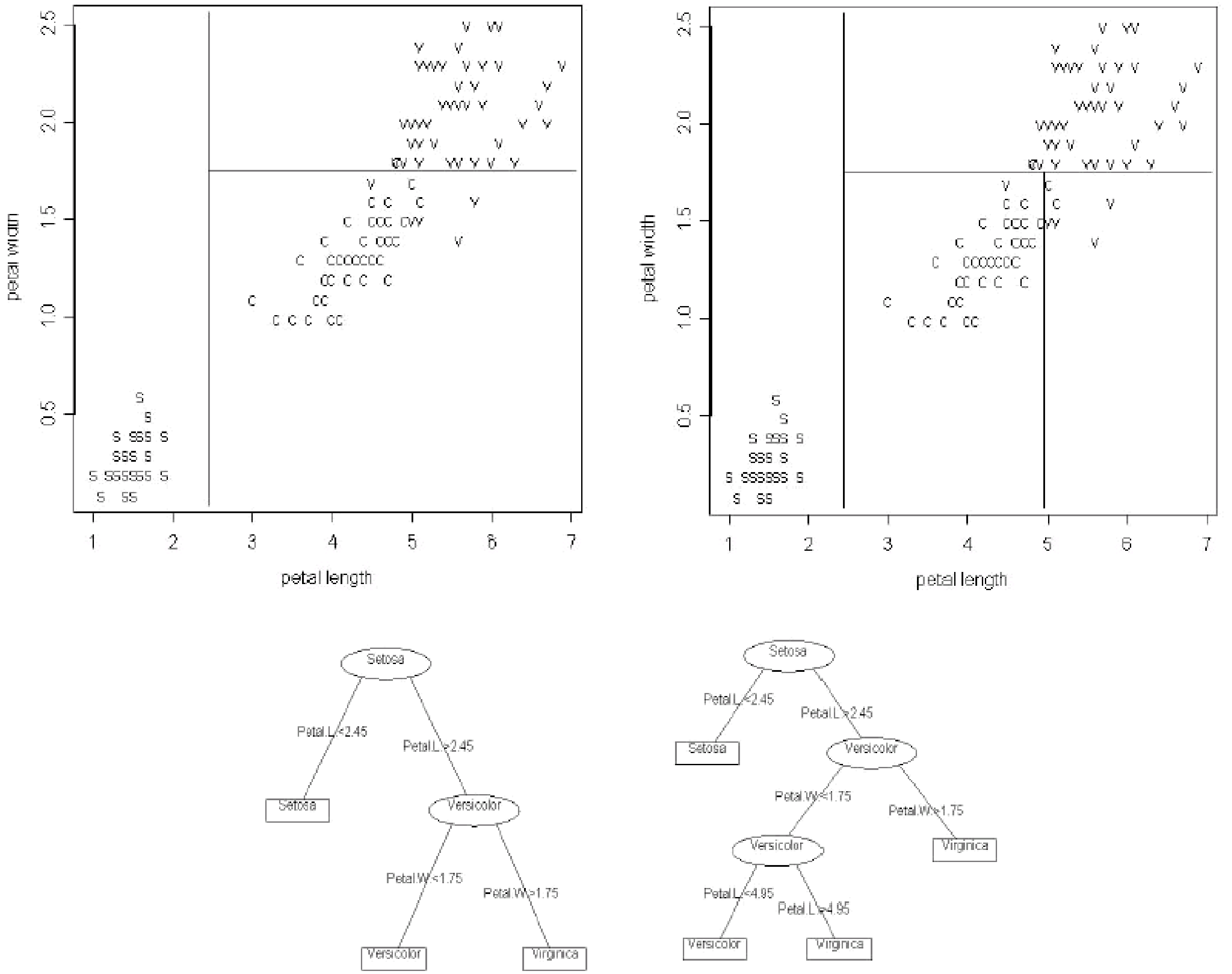






Growing

Growing Tree (4)



Conventional criteria – Gini index

- The class labels are denoted 0 and 1.
- Given a split into left and right buckets, let $p_L^0 + p_L^1 = 1,\ p_R^0 + p_L^1 = 1$
- The Gini index is $p_L^0 p_L^1$ in left bucket.
- This is essentially the Mean Square Error (MSE) when fitting a mean to $y_n \in \{0,1\}, n \in L$, as a short calculation shows:

$$
\begin{aligned}
\min MSE_L &= \min \frac{RSS}{N_L} = \min_q \frac{1}{N_L} \sum_{n\in L} (y_n - q)^2 \\
&= \frac{1}{N_L} \sum_{n\in L} (y_n - p_L^1)^2 \\
&= \frac{1}{N_L} (N_L^0 (p_L^1)^2 + N_L^1 (1 - p_L^1)^2) \\
&= \frac{1}{N_L} (N_L p_L^0 (p_L^1)^2 + N_L p_L^1 (p_L^0)^2) \\
&= \frac{1}{N_L} N_L p_L^0 p_L^1 (p_L^1 + p_L^0) = p_L^0 p_L^1
\end{aligned}
$$

Pruning in CART

- To avoid shortness of trees, a large tree T_0 is grown and then pruned backward.
- Pruning criterion: cost of a subtree $T \in T_0$, defined by

$$C_\alpha(T) = \Sigma D_g(T) + \alpha \cdot |T|$$

- Here the sum is over the terminal nodes of T, $|T|$ is the number of terminal node in T and α is a cost-complexity parameter.

Pruning in CART (cont.)

- For each fixed α, the best subtree T_α is found via weakest link pruning.
- Large α gives smaller trees.
- A best value $\hat{\alpha}$ is estimated via 10-fold cross-validation.
- Final chosen tree is $\widehat{T}_{\hat{\alpha}}$.

What is “surrogate”?

- Meaning : to put in the place of another, to appoint as successor
- Once a splitting variable and a split point for it have been decided, what is to be done with observations missing that variable?
- One approach is to estimate the missing datum using the other independent variables, that is *surrogate variables*.

Example of “surrogate”

- Assume the split (age < 40, age >= 40) has been chosen.
- The surrogate variables are found by
  - Re-applying the splitting algorithm (without recursion)
  - To predict the two categories “age < 40, age >= 40”
  - Using the other independent variables.
- Any observation which is missing the split variable is then classified using the first surrogate variable,
  - Or if missing that, the second surrogate is used.
  - and etc.

Tree regression: why?

- Tree regression had not been popular due to inaccuracy.
- But tree regression is very attractive since the model is interpretable.
- Friedman, Hastie, and Tibshirani (2000) showed that tree regression can be used in classification problem.

Regression

Tree regression: 4 ingredients

- Splitting criterion

$$SS_T - (SS_L + SS_R)$$

  - Where $SS_T = \Sigma(y_i - \bar{y})^2$ is the sum of squares of the node, and SS_L, SS_R are the sums of squares for the right and left son, respectively.
  - Equivalently to choosing the split to maximize the between-groups sum-of-squares in a simple analysis of variance
- A summary statistics (describing a node)
  - For the anova method the response is the mean of the node
  - For classification the predicted class followed by the vector probabilities
- The error of a node : the variance of y for anova
- The prediction error for a new observation : $(y_{new} - \bar{y})$

S-PLUS

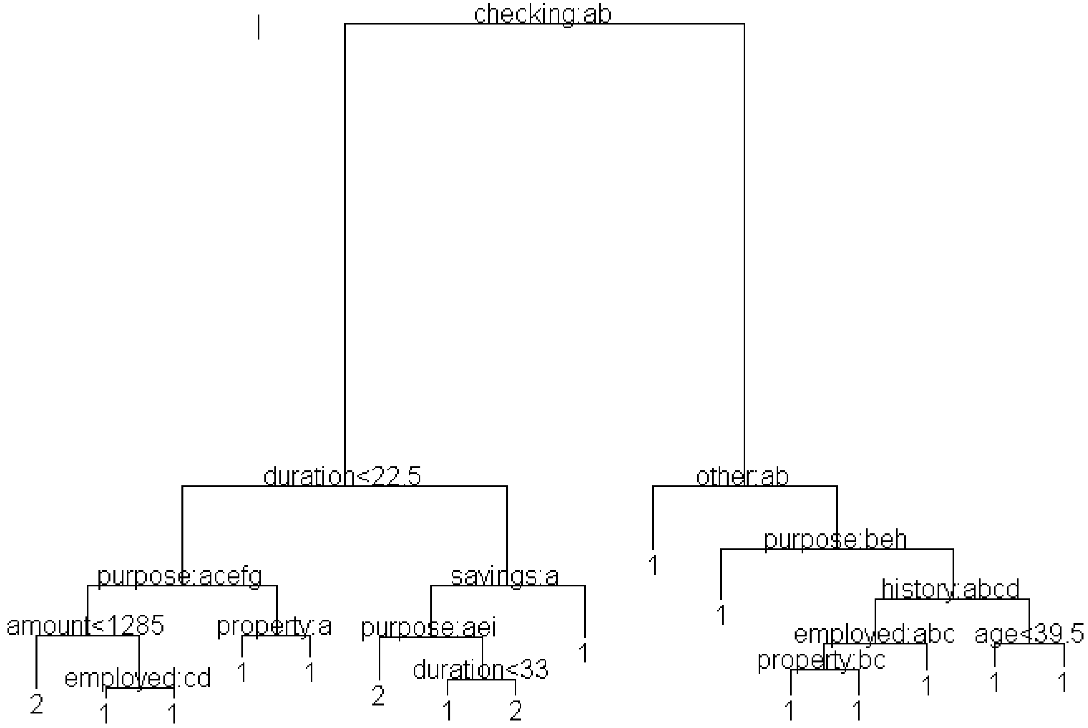
German Credit Data

- Observation 1,000개
- 20개의 설명변수 중, 7개는 연속형, 나머지는 이산형
- Target variable은 credit의 good과 bad.
- 주요 설명변수
  - checking: status of existing checking account (4 levels)
  - savings: savings account/bond (5 levels)
  - duration: duration in month
  - employment: present employment since (5 levels)
  - ……
- http://www.ics.uci.edu/AI/ML/Machine-Learning.html

S-PLUS

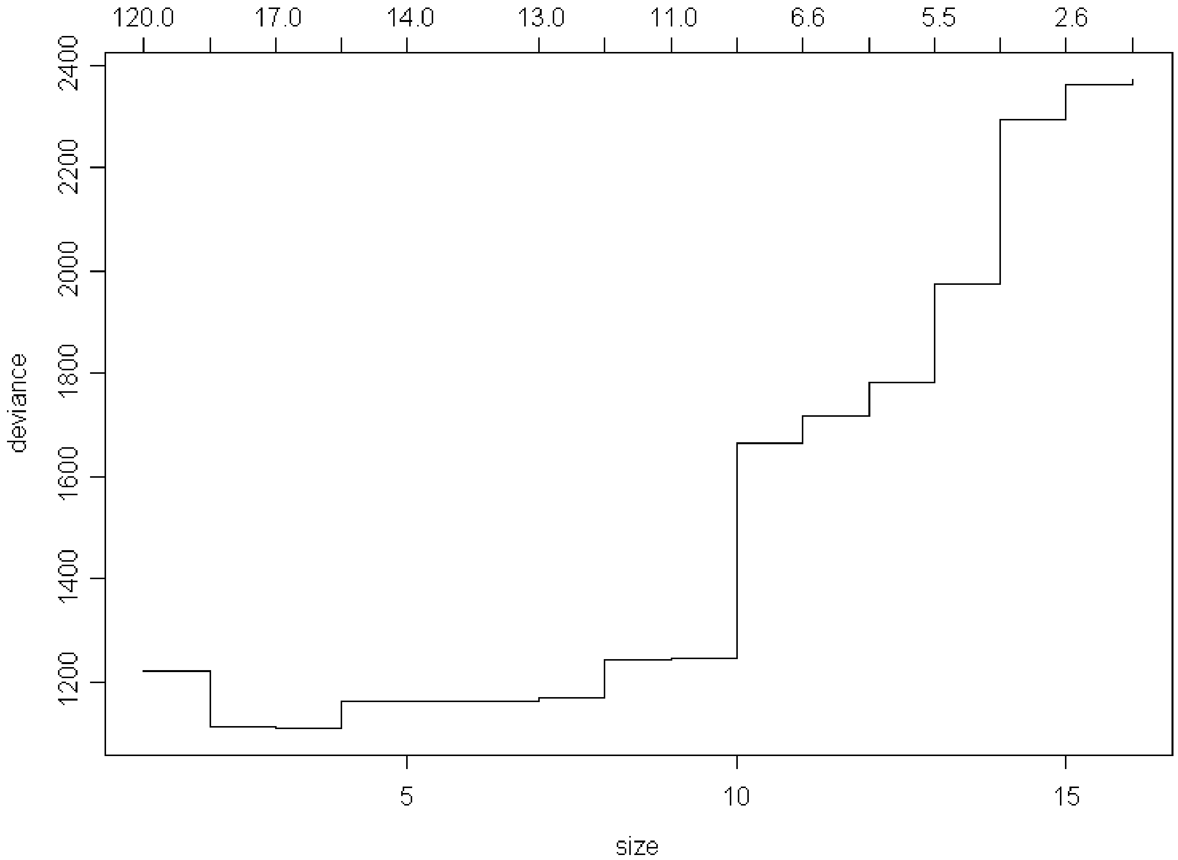
Tree by deviance

```
> german.tree <- tree(good.bad~., data=german)
> plot(german.tree)
```



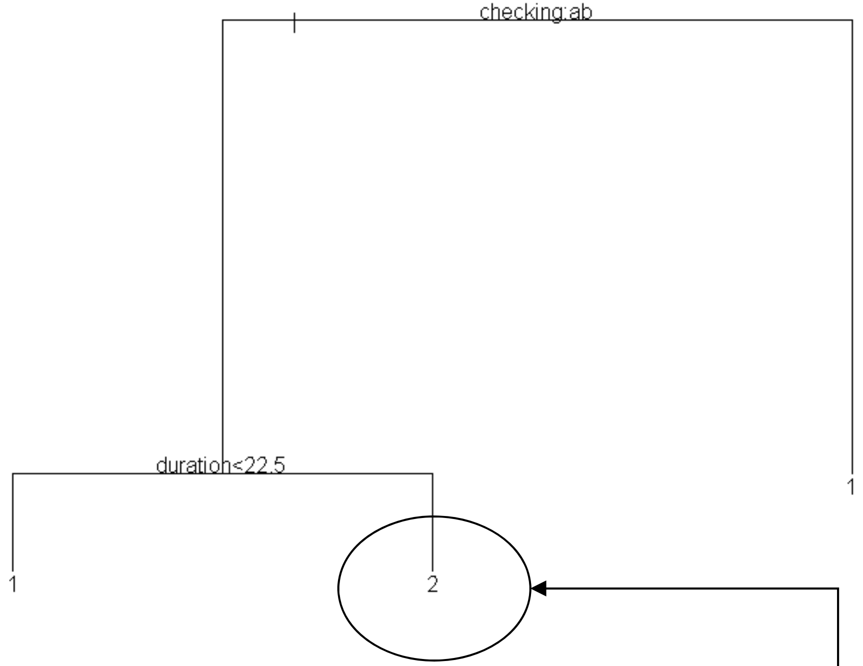
Cross-validation

```
> german.cv <- cv.tree(german.tree,FUN=prune)
> plot(german.cv)
```



Tree by deviance (after pruning)

```
> german.tree.3 <- prune.tree(german.tree, best=3)
> plot(german.tree.3)
```



Bad credit의 비율: 0.5646
No. of obs.: 247개
Rule: checking-1,2/ duration>22.5

Q&A