Tree-based methods II

Roberta De Vito

BROWN
Public Health

Classification Trees

- Prediction: each observation belongs to the most commonly occurring class
- Not only class prediction but also class proportion
- Classification error rate

$$E = 1 - max_k(\hat{p}_{mk})$$

Impurity measure: Gini Index (Q1)

- The Gini Index

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Impurity measure: Entropy measure (Q2)

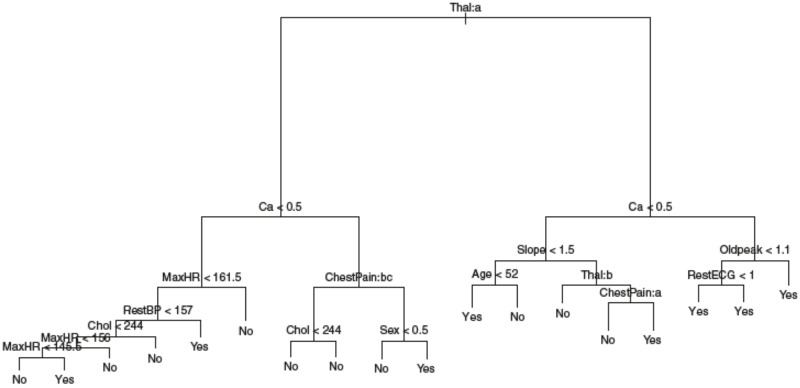
- The entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk})$$

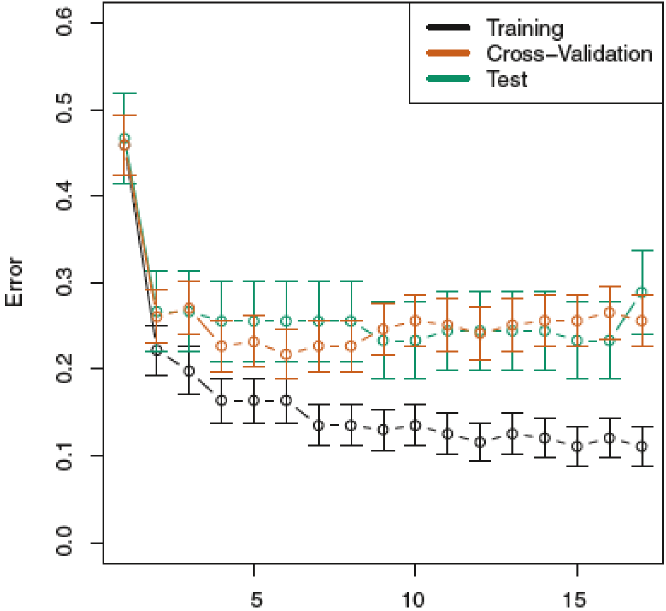
Example with Heart Data set

- 303 patients with chest pain
- Y: 0 (no heart disease), 1(yes)
- 13 predictors: Age, Sex, Chol

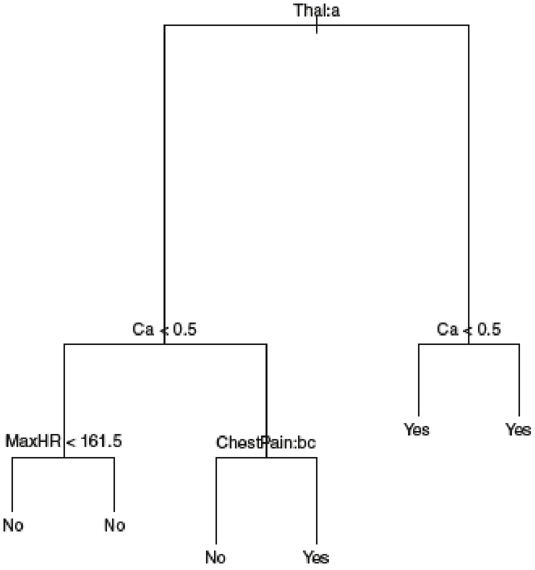
A look to the tree (Q3)



The cross-validation: Q4



The pruned tree



Tree vs Linear Models

- Linear regression

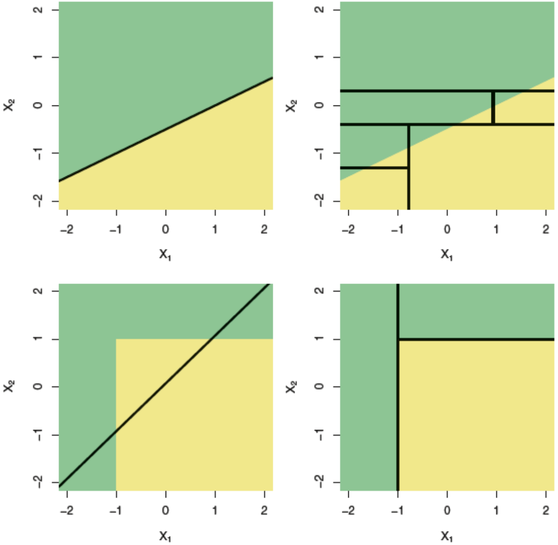
$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- Regression tree

$$f(X) = \sum_{m=1}^{M} c_m 1_{X \in R_m}$$

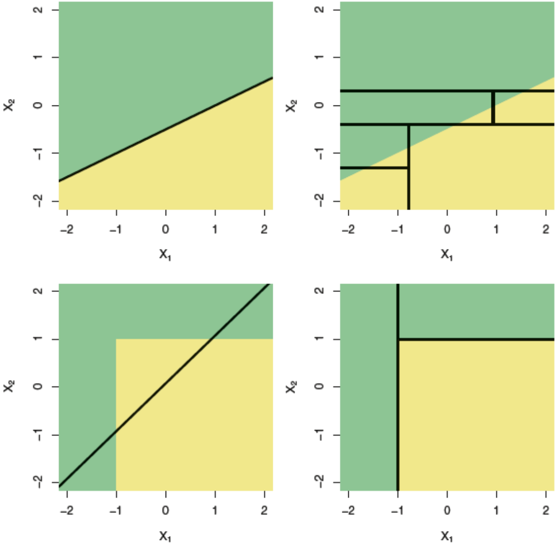
Tree vs Linear Models

Q5



Tree vs Linear Models

Q6



Tree vs Linear Models

- Easy to understand and visualization
- Closely mimic human decision
- No need to create a dummy

Tree vs Linear Models

- Easy to understand and visualization
- Closely mimic human decision
- No need to create a dummy

- Prediction
- Robustness

Bagging

- Bootstrap

- We could calculate $\hat{f}^1(x), \ldots, \hat{f}^B(x)$:

$$\hat{f}_{bag}(x) = \frac{1}{B}\hat{f}^b(x)$$

- Improvement in accuracy

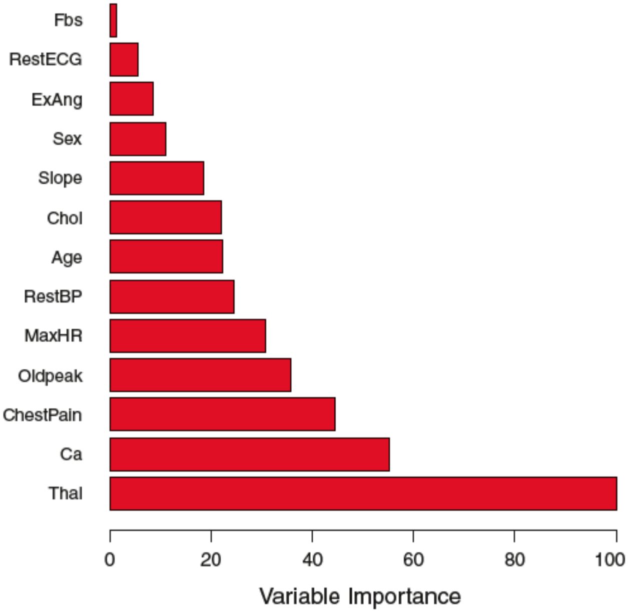
Out-of-Bag Error Estimation

- Each bagged tree makes use of around two-thirds of the observations
- The remaining one-third: out-of-bag (OOB)
- Prediction: using the OOB
- $B/3$ predictions: average
- B sufficiently large $\rightarrow$ OOB error is equivalent to leave-one-out CV

Variable Importance Measures

- It is no longer clear which variables are most important to the procedure
- The total amount that the RSS is decreased due to splits over a given predictor, averaged over all B trees
- A large value indicates an important predictor

Variable Importance Measures



Random Forests

- Building decision trees $\rightarrow$ we consider a random sample of m predictors
- $m \approx \sqrt{p}$
- The trees will use this strong predictor $\rightarrow$ predictions are highly correlated
- Random Forest: $(p - m)/p$ of the splits will not consider strong predictors

Test and OOB error