Tree-based methods

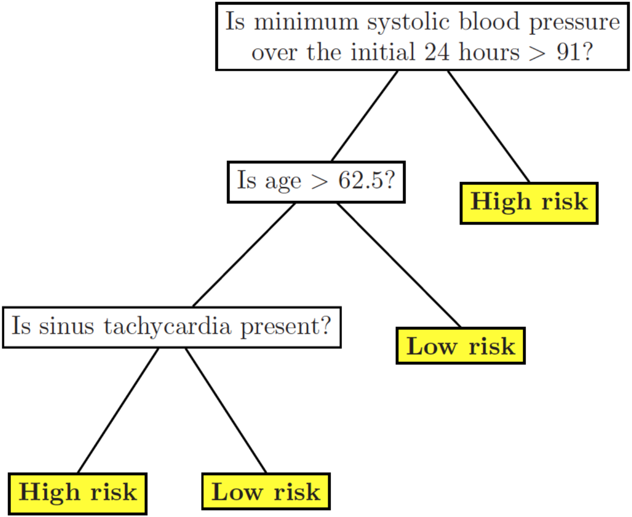
Roberta De Vito

BROWN
Public Health

General Idea

- Classify observations into known classes
- Predict levels of regression functions
- Decision Tree
- Improvements in prediction accuracy
- Non-parametrized method: Q1 in prismia

How do the tree-based methods work ? Q2 prismia



Is minimum systolic blood pressure over the initial 24 hours > 91?

Is age > 62.5?

High risk

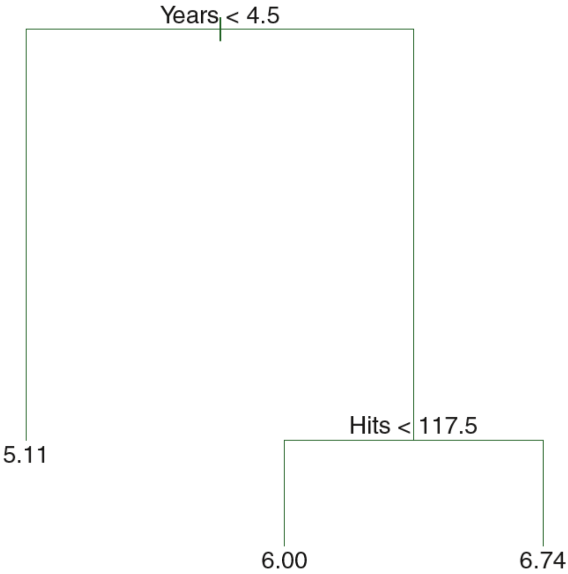
Is sinus tachycardia present?

Low risk

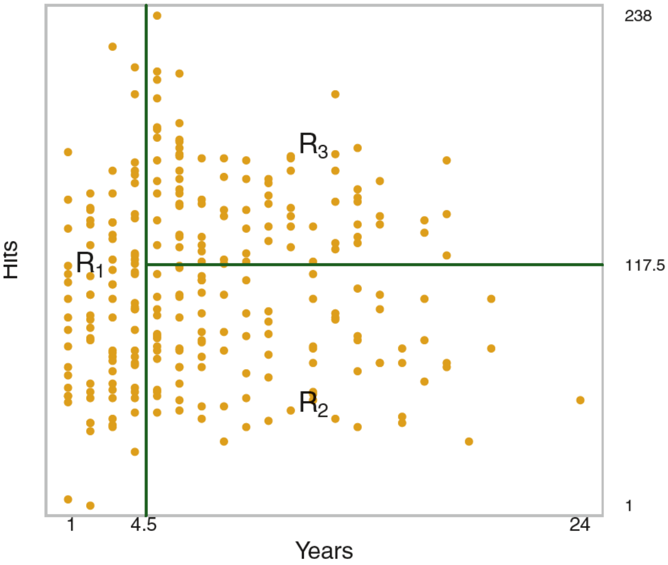
High risk

Low risk

How do the tree-based methods work ? Q3-Q4 prismia



The regions



Process of building a regression tree

- Divide the predictor space $X_1, X_2, \ldots, X_p$ into J regions $R_1, \ldots, R_j$
- For every observation that falls into the region $R_j \rightarrow$ same prediction: Q6 prismia

How do we construct the regions $R_1, ..., R_J$?

- Rectangular boxes but any shape
- Minimize (Q7 prismia)

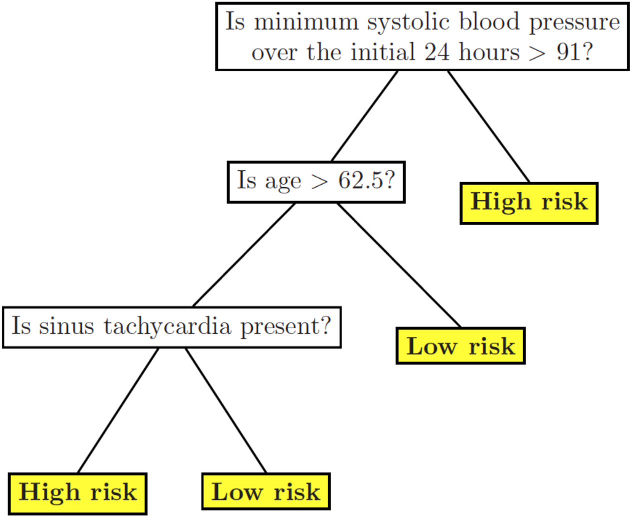
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- Consider all predictors

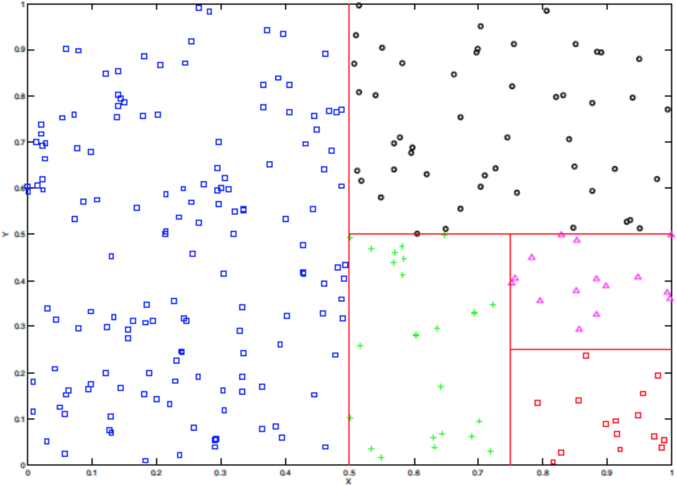
$$R_1(j,s) = \{X | X_j < s\} \text{ and } R_2(j,s) = \{X | X_j \geq s\}$$

- Look for the third region: split one of the two previously identified regions

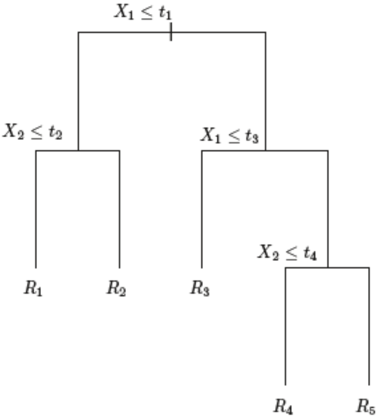
Example: Q8

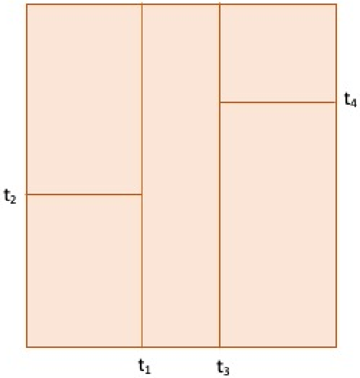

Is minimum systolic blood pressure over the initial 24 hours > 91?
Is age > 62.5?
High risk
Is sinus tachycardia present?
Low risk
High risk
Low risk

Example: Q8

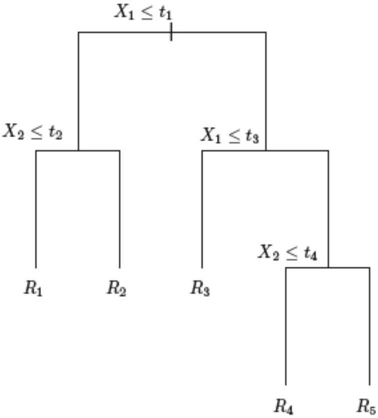


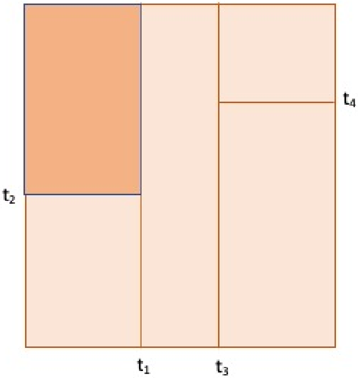
Example with questions



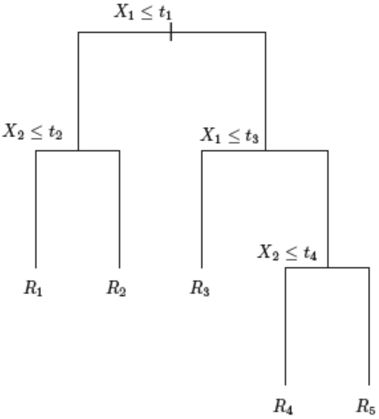


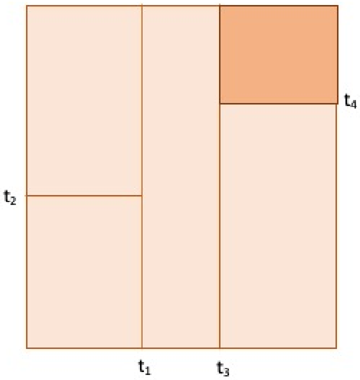
Example with questions Q9



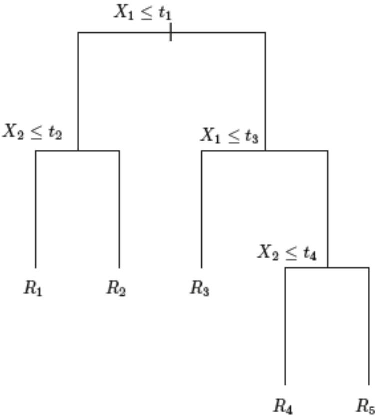


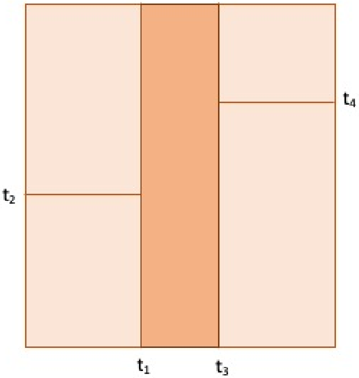
Example with questions: Q10



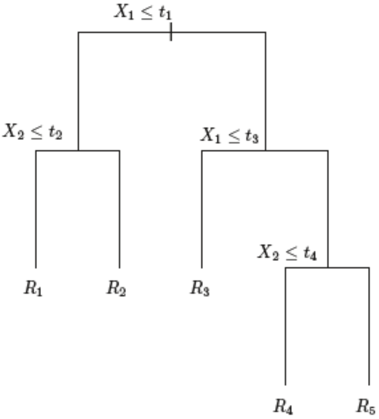


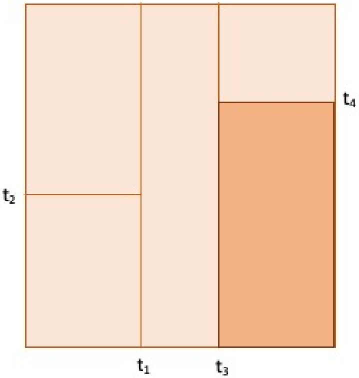
Example with questions



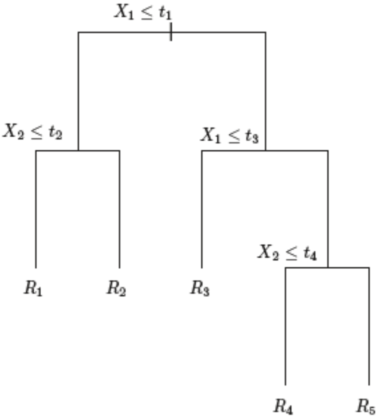


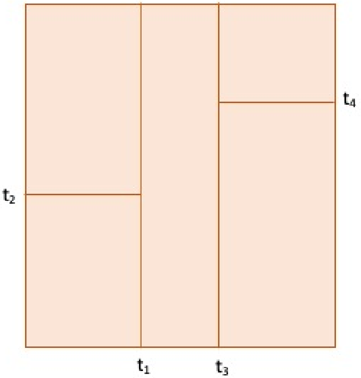
Example with questions





Example with questions





Tree Pruning

- Good prediction on training set, poor test performance
- Build the tree: decrease of RSS exceeds a threshold
- Better strategy: grow a very large tree T_0, and then prune it back in order to obtain a subtree
- How do we determine the best prune way to prune the tree?
- Test error using cross validation

Cost Complexity Pruning

- Consider a sequence of trees indexed by a nonnegative tuning parameter α
- For each α: $T \subset T_0$

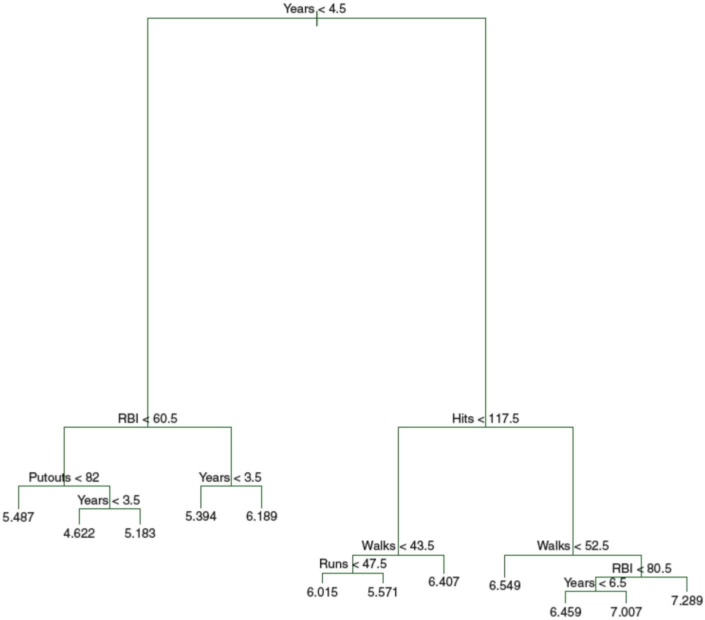
$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

- $\alpha = 0 \rightarrow T = T_0$
- α increase $\rightarrow$ smaller subtree
- How can we select α?

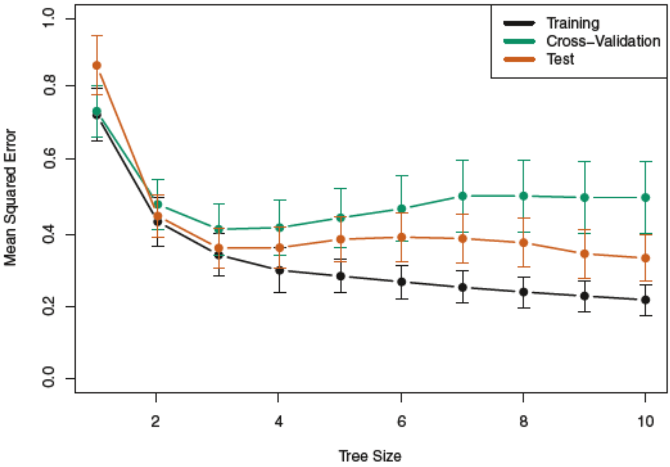
Data: hitters

- Randomly divided the data set in half (132 in the training set, and 131 in the test set)
- A large regression tree on the training data and varied α
- Six-fold cross-validation: estimate the cross-validated MSE

Data: hitters



Data: hitters


Training
Cross–Validation
Test
Mean Squared Error
Tree Size