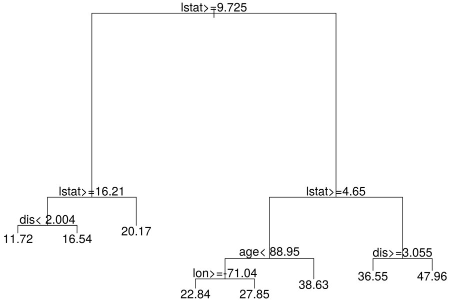
Decision Trees I

Tree Pruning

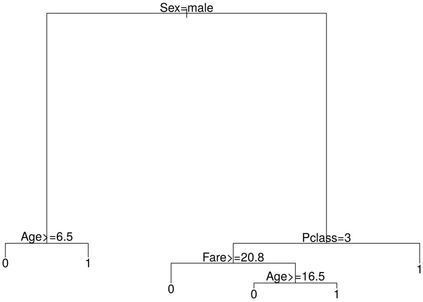
Tree structure

Figure: CART examples

(a) Regression tree

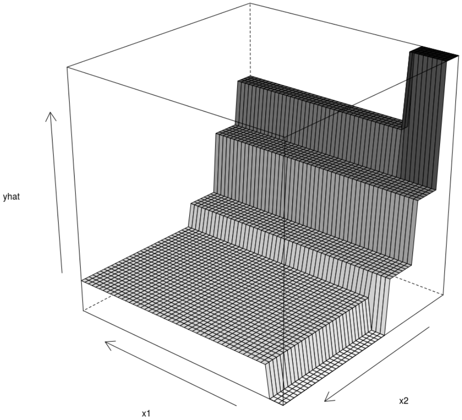


(b) Classification tree



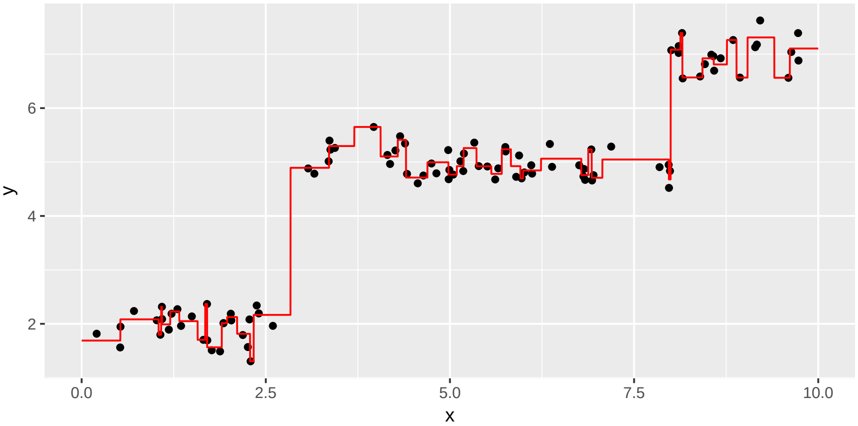
Tree structure

Figure: Tree prediction surface (example)



Tree structure

Figure: High variance in trees



- Overfitting = Poor generalization to new data
- Function approximates training data well, but the number of terminal nodes is high

Tree pruning

Stopping rules

- Minimum number of cases in terminal nodes
- Decrease in impurity exceeds some threshold

$\rightarrow$ However, worthless splits can be followed by good splits

Cost complexity pruning

Find optimal subtree(s) $\mathcal{T}_\alpha$ by balancing tree quality $SSE(\mathcal{T}) = \sum(y_i - \hat{y}_i(\mathcal{T}))^2$ and tree size $|\mathcal{T}|$

$$C_\alpha(\mathcal{T}) = SSE(\mathcal{T}) + \alpha|\mathcal{T}|$$

- α controls the penalty on the number of terminal nodes
- α can be chosen through CV

Tree pruning

Stopping rules

- Minimum number of cases in terminal nodes
- Decrease in impurity exceeds some threshold

$\rightarrow$ However, worthless splits can be followed by good splits

Cost complexity pruning

Find optimal subtree(s) $\mathcal{T}_\alpha$ by balancing tree quality $SSE(\mathcal{T}) = \sum(y_i - \hat{y}_i(\mathcal{T}))^2$ and tree size $|\mathcal{T}|$

$$C_\alpha(\mathcal{T}) = SSE(\mathcal{T}) + \alpha|\mathcal{T}|$$

- α controls the penalty on the number of terminal nodes
- α can be chosen through CV

Surrogate splits and costs

Missings

- Create a new category for missing values
- Use surrogate splits
  1. Choose best (primary) predictor based on complete cases
  2. Search for surrogate variables which mimic the chosen split
  3. Use surrogates if values for primary predictor are missing

Costs

$$\mathbf{L} = \begin{pmatrix} 0 & L_{fp} \\ L_{fn} & 0 \end{pmatrix}$$

- Typically $L_{fp} = L_{fn} = 1$
- Misclassifications can be weighted differently
  - Modification of loss-matrix through weights / modified Gini index

Surrogate splits and costs

Missings

- Create a new category for missing values
- Use surrogate splits
  1. Choose best (primary) predictor based on complete cases
  2. Search for surrogate variables which mimic the chosen split
  3. Use surrogates if values for primary predictor are missing

Costs

$$\mathbf{L} = \begin{pmatrix} 0 & L_{fp} \\ L_{fn} & 0 \end{pmatrix}$$

- Typically $L_{fp} = L_{fn} = 1$
- Misclassifications can be weighted differently
  - Modification of loss-matrix through weights / modified Gini index

Summary

- Divide-and-conquer strategy that splits the data into subgroups
- Surface from decision trees is a non-smooth step function
- No need to specify the functional form in advance (unlike regression)
- Non-linearities and interactions are handled automatically
- Limitations: Instability(!), competition among correlated predictors, biased variable selection

Software Resources

Resources for R

- Basic CART implementation: `tree`
- Standard package to build CARTs: `rpart`
  - Includes build-in Cross-Validation and `prune` function
- Unified infrastructure for tree representation: `partykit`

References

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review 82*(3), 329–348.

Zhang, H., Singer, B. (2010). *Recursive Partitioning and Applications*. New York, NY: Springer.