Regression Trees

Timothy Currie

Wissenschaftliches Arbeiten

25/06/2024

Supervisor: Dr. Elias Wolf

Matrikelnummer: 50074426

Universität Bonn

- Linear regression performs poorly on many kinds of data.
- E.g. data with non-linear relationships and interaction effects.
- What might be a better approach for these situations?
- Regression Trees can be useful in many situations.

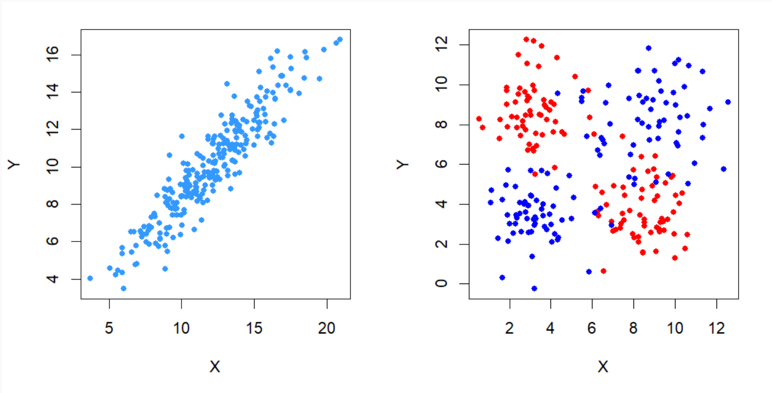


Figure 1: Can you guess where regression trees and where linear regression will perform better?

Outline of the Talk

1. Motivation
2. Basics of Regression Trees
3. Comparing regression Trees and Linear Regression
4. Overfitting and Pruning
5. Ensemble methods and BART
6. Conclusion
7. References

Basics of Regression Trees

- Regression trees split the predictor space into regions that minimize the RSS given by

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

- In each region $\hat{y}$ simply takes on the mean of all observations in that region.
- Typically we can't find the optimal regions.
- Instead we use a greedy algorithm **Recursive binary splitting** to find the optimal split to minimize prediction error at each stage.

- **Main differences between linear regression and trees**
  - Regression trees do not assume a linear relationship between predictors and the response.
  - Trees capture interaction effects naturally.

Simulation: Linear Data

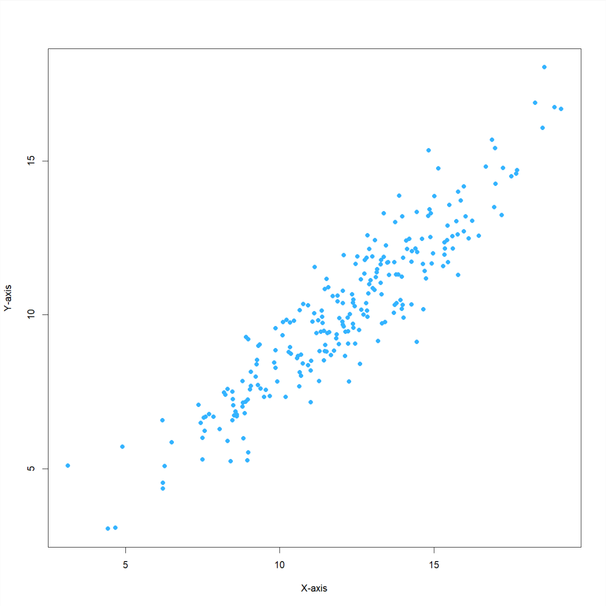


Figure 2: Linear Relation between Variables

- We will run all simulations 400 times.
- For now the regression trees will have 4 terminal Nodes.
- And we will always compare the Mean Squared Error (MSE).
- Data generated as: $Y = \beta_0 + \beta_1 X + \epsilon$
- Where $\epsilon \sim N(0, \sigma^2)$

Simulation: Linear Data

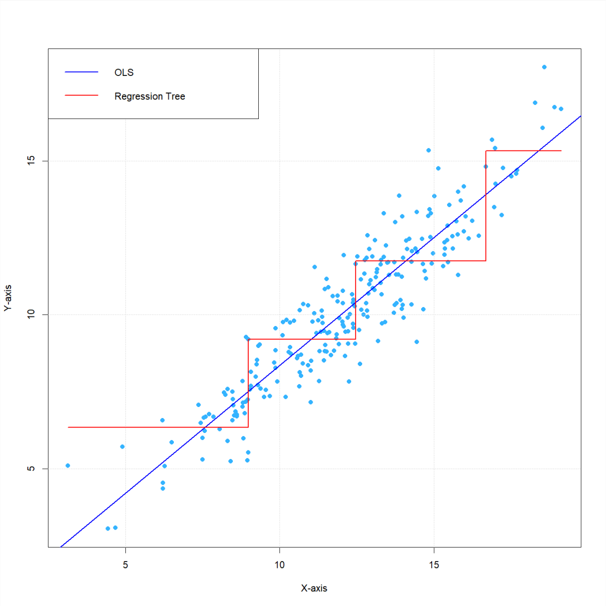


Figure 3: Linear Relation between Variables

- We will run all simulations 400 times.
- For now the regression trees will have 4 terminal Nodes.
- And we will always compare the Mean Squared Error (MSE).

- **Results:**
- **MSE** for OLS model: **0.9917**
- **MSE** for regression tree model: **1.4935**

Simulation: Non-linear Data

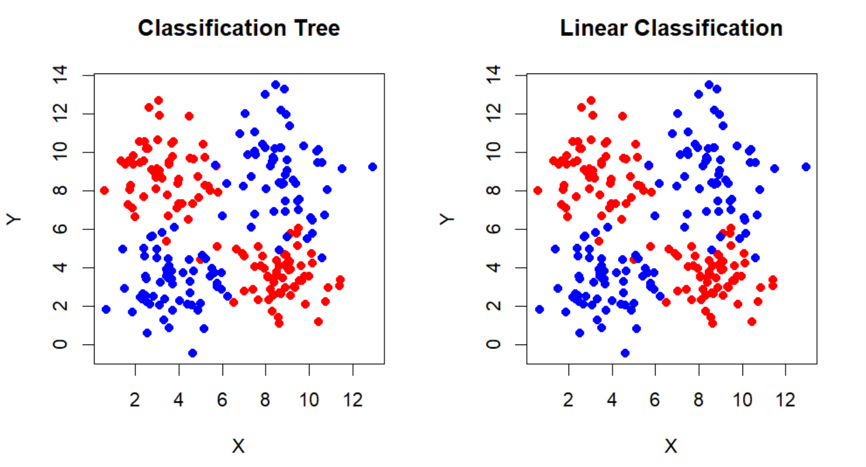


Figure 4: Data with non-linear relationship

- Non-Linear Data generated using 4 normal distributions
- NW & SE = red, NE & SW = blue

Simulation: Non-linear Data

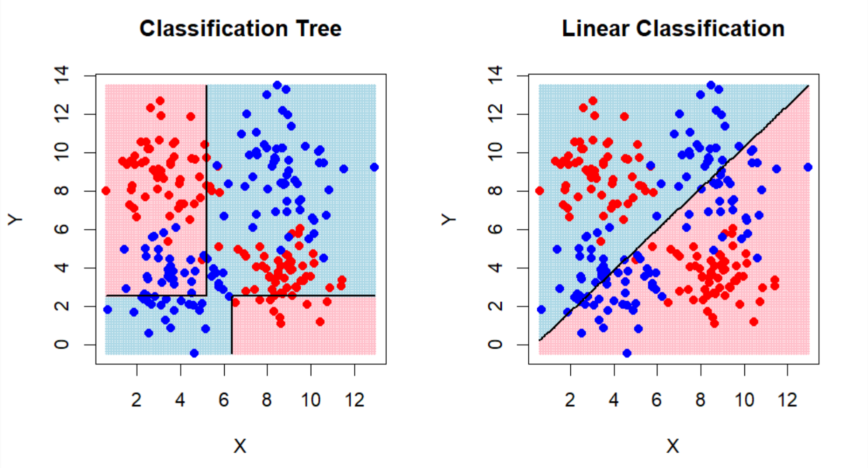
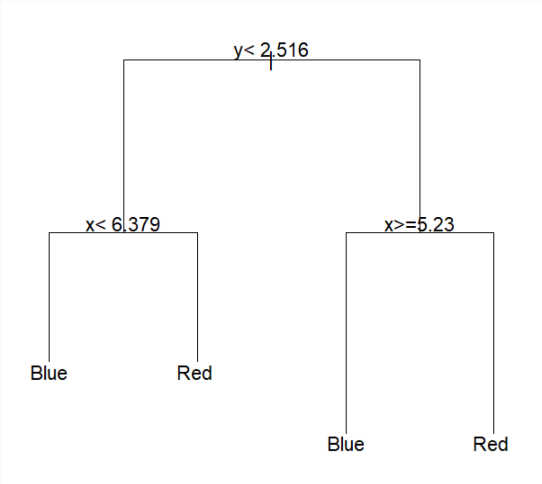


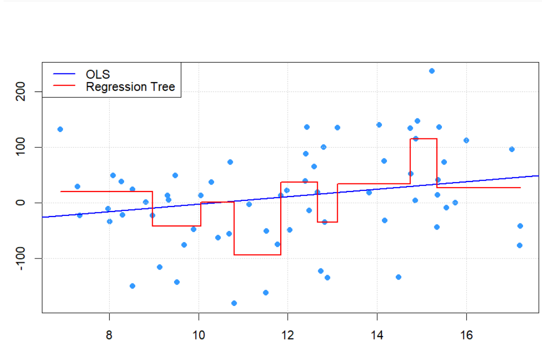
Figure 5: Data with Prediction

- Classification Tree MSE: 0.3757
- Linear Classification MSE: 0.5103
- **Trees capture interaction:** e.g. Large Y is only indicative of red if X is small.

Two Problems with Regression Trees



(a) Greedy Classification Tree



(b) Better Overfitting

Figure 6: Two Problems that can arise with Trees

- Overfitting and non-Optimal Splitting

Pruning

- **Cost complexity pruning** counteracts overfitting by removing non-essential splits.
- Lets us grow a large Tree and then

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

  - Select a parameter α.
  - For each α, find the subtree that minimizes the cost.
  - Use cross-validation to select the best α.
  - Instead of evaluating a Model on the data we trained it on we evaluate it on a separate set.
- **Objective:** Achieve a good tradeoff between bias and variance.
- Large α results in very small trees, small α in larger trees.

Simulation: Finding the optimal Tree size

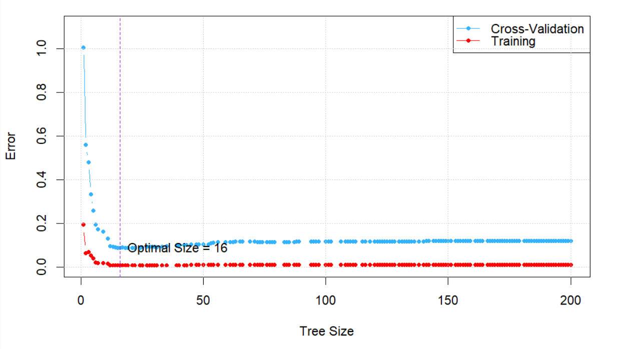


Figure 7: Training and Cross-Validation Error

- Pruning helps against overfitting and improves overall performance
- Original Tree MSE on Test set = 198.2861
- Pruned Tree MSE on Test set = 173.861
- Sweet spot balances variance and bias, minimizing overfitting.
- Training set used to build model, test set to evaluate performance

Ensemble methods

- Even with pruning trees often perform worse than linear other ML methods
- Ensemble methods improve results by combining many regression trees. Each one contributes a small part to the overall prediction.
- Each tree can be independent of previous trees (e.g. Random Forests)
- Or can be grown on the residuals of the current fit (e.g. Bayesian Additive Regression Trees (BART))

BART

- BART models the response as a sum of many tree-based models plus noise.
- **Model:**

$$Y_i = \sum_{j=1}^{m} g(X_i; T_j, M_j) + \epsilon_i. \tag{1}$$

- BART calculates the residuals of the current sum of Trees.
- Then modifies one Tree to decrease the residuals.
- Then take the average over all but the burn-in iterations.
- Unlike single trees, BART avoids overfitting by averaging the predictions of many trees.
- BART provides a probabilistic prediction, giving a measure of uncertainty.

Conclusion and Discussion

- Regression trees are powerful for non-linear and interactive effects.
  - Will often outperform linear regression.
- They are also very easy to interpret.
- Trees require pruning to combat overfitting.
- By averaging independent Trees or fitting trees on the residuals ensemble methods can improve results.
- BART is a sophisticated method offering good results in many scenarios.

References

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (Second Edition). Springer.
- Tan, Y. V., Roy, J. (2019). Bayesian additive regression trees and the General BART model. Statistics in Medicine, Band/Volume 38(25), 5048-5069.
- Chipman, H. A., George, E. I., McCulloch, R. E. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, Band/Volume 4(1), 266-298.
- Townshend, R. Lecture 10 - Decision Trees and Ensemble Methods — Stanford CS229: Machine Learning (Autumn 2018). https://www.youtube.com/watch?v=wr9gUr-eWdA, accessed 21.05.24.

- All images were made using R.
- Also thanks to Claude and ChatGPT for making LaTeX a lot nicer to use.