RECSM Summer School: Machine Learning for Social Sciences

Session 2.2: Advantages and Disadvantages of Trees

Reto Wüest

Department of Political Science and International Relations
University of Geneva

Outline

1. Trees Versus Linear Models

2. Advantages and Disadvantages of Trees

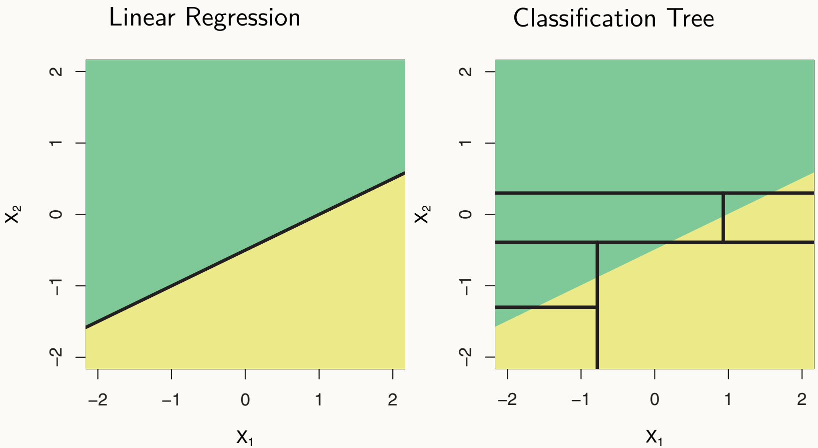
Trees Versus Linear Models

Trees Versus Linear Models

- When does a regression tree perform better than linear regression?
- If the relationship between the predictors and the response is well approximated by a linear model, then linear regression will outperform a method such as regression tree that does not exploit this linear structure.
- If the relationship between the predictors and the response is highly non-linear and complex, then a regression tree may outperform linear regression.
- Note that the relative performances of regression tree and linear regression can be assessed by estimating the test error, e.g., using CV.

Trees Versus Linear Models: Example

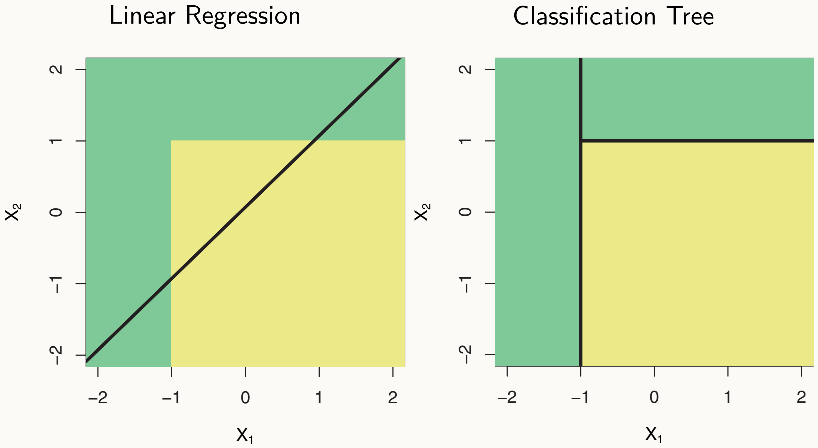
Two-Dimensional Classification Problem With a Linear Decision Boundary



(Source: James et al. 2013, 315)

Trees Versus Linear Models: Example

Two-Dimensional Classification Problem With a Non-Linear Decision Boundary



(Source: James et al. 2013, 315)

Advantages and Disadvantages of Trees

Advantages of Trees

- Trees are very easy to explain to people. In fact, they are even easier to explain than linear regression.
- Decision trees might mirror human decision-making more closely than do the classical regression and classification approaches.
- Trees can be displayed graphically, and are easily interpreted even by a non-expert (especially if they are small).
- Trees can easily handle qualitative predictors without the need to create dummy variables.

Disadvantages of Trees

- In general, trees do not have the same level of predictive accuracy as other supervised learning methods (e.g., shrinkage methods).
- Trees can be very non-robust: a small change in the data can cause a large change in the final estimated tree.

⇒ By aggregating many decision trees (bagging, random forests, boosting), the predictive performance of trees can be substantially improved.

⇒ Bagging, random forests, and boosting use trees as building blocks to construct more powerful prediction models.