Lecture 9 Decision Trees

wbg231

December 2022

introduction

- Decision trees are an inherently non linear type of model
- they are lso a good way to understand Ensemble methods.

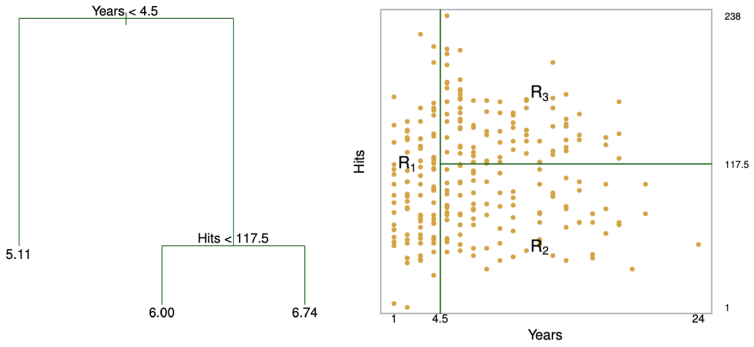
Decision trees

- regression trees try to predict a continuous outcome
- classification trees try to predict a discrete class.
- a binary tree has 2 children nodes there are also multi-way trees that have more
- each node contains a subset of teh data
- the data splits created by each node involve only a single feature
- for continuous variables the splits are always of the form $x_i \leq y$
- for discrete variables we partition the values into two stets
- predictors are made in the leaf nodes

constructing the tree

- our goal is to find the boxes $R_1 \dots R_J$ that minimize $\Sigma_{j=1}^{J} \Sigma_{i \in R_j} (y_i - \hat{y}_{R_j})$ subject to complexity constraints
- the issue with this is finding the true optimal binary tree is computationally intractable
- so instead we use a greedy algorithm (that is we take the best choice at every step) where we start at the root and on our first step take the splits that would result in the minimal loss, and then pass the data split like that to the next step and have each for each of those sections pick the best splits and continue this until we hit some stopping criteria

- we only split regions defined in the last step at each current step we predict based on the mean value of a terminal node ie $\hat{y}_{R_m} = mean(y_i | x_i \in R_m)$
- so building a tree like this we are making the best local choice at every step but are unlikely to reach the overall optimal choice



-
-