Decision Tree

Caleb Jin

9-27-2019

Contents

Foreword

I am Caleb Jin. This is my note of **decision tree**. When I try to grasp a statistical method, I'd like to wirte down every details about it that I am able to. I mainly use **An Introduction to Statistical Learning with Applications in R** (James et al., 2014) and **The Elements of Statistical Learning** (Hastie et al., 2001). Due to my limited statistics knowledge, if making any mistakes, I sincerely expect you guys can email to me. My email address is jinsq@ksu.edu. Appreicate!

Chapter 1

The Basic of Decision Tree

Let's start with a simple model setting. Consider we have a continuous response variable $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and 2 predictors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{n\times 2}$.

The decision tree starts with splitting a predictor, say $\mathbf{x}_1$,

- 1) We partition $\mathbf{x}_1$ into two distinct regions $R_1(1,s) = \{\mathbf{x}_1 | \mathbf{x}_1 < s\}$ and $R_2(1,s) = \{\mathbf{x}_1 | \mathbf{x}_1 \geq s\}$.
- 2) As all observations are divided into two regions $R_1(1,s)$ or $R_2(1,s)$, then we make the same prediction with

$$\hat{y}_{R_1} = \frac{1}{n_1} \sum_{i:x_{i1}\in R_1} y_i,$$

$$\hat{y}_{R_2} = \frac{1}{n_2} \sum_{i:x_{i1}\in R_2} y_i.$$

The question is how to determine s.

From the step 1), we get $R_1(1,s)$ and $R_2(1,s)$ by spliting $\mathbf{x}_1$, for example. We hope this splitting can maximize sum of squares between regions and minimize sum of squares within regions of $\mathbf{y}$. As total sum of squares of $\mathbf{y}$ is fixed, maximium of sum of squares between regions is equivalent to minimum of sum of squares within regions. This lead us to consider a classic criterion, residual sum of squares (RSS):

$$RSS = \sum_{j=1}^{2} \sum_{i:x_{i1}\in R_j(1,s)} (y_i - \hat{y}_{R_j})^2.$$

Bibliography

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.