CLASSIFICATION TREES

Data Mining

Cesar Acosta

February 19, 2020

INTRODUCTION

- A model to predict a categorical response with k levels
- Example
  - Predict the gender of a new customer
  - based on customer attributes (predictors)
  - attributes may be continuous or categorical

INTRODUCTION

- A model to predict a categorical response with k levels
- How?
  - Divide the predictors space into many regions
  - In each region there are observations from the k classes (levels)
  - Find p_{mk} the proportion of observations from the k class in the m^{th} region
  - For each region m, the prediction $\hat{y}_m$ is the most common class
  - For each region, the error rate is the fraction of obs that do not belong to the most common class

INTRODUCTION

- A model to predict a categorical response with k levels
- How?
- Example - Consider a response with $k = 3$ classes
  - For region $m = 4$ $\begin{cases} p_{41} = 10\% & \text{members from class 1} \\ p_{42} = 20\% & \text{class 2} \\ p_{43} = 70\% & \text{class 3} \end{cases}$
  - Prediction is $\hat{y}_4 = 3$
  - error rate for regions 4 is $e_4 = 0.3$
  - region 4 would be *pure* if $p_{4j} = 1$ for some class j

INTRODUCTION

- Different measures of *purity*
  - E: classification error rate $$E = \sum_{m=1}^{T} e_m$$
  - G: Gini Index $$G = \sum_{m=1}^{T} \sum_{i=1}^{K} p_{im} \left(1 - p_{im}\right)$$
  - D: Cross entropy $$D = -\sum_{m=1}^{T} \sum_{i=1}^{K} p_{im} \ln \left(p_{im}\right)$$

CROSS VALIDATION

- Fit trees of different depths using the training set
- Find their training error rates
- Select depth of the tree with the smallest training error rate
- Prune a full tree to the selected depth
- Find the *test* error rate of the pruned tree

CROSS VALIDATION

- Use function `cv.tree`
- `cv.tree(tree1)` compares regression trees based on *deviance*
- Use `cv.tree(tree1, FUN=prune.misclass)`
  to compare categorical trees based on the number of misclassified observations

CROSS VALIDATION

- The MSE or RSS of a tree with two regions R_1 and R_2 is

$$RSS = \sum_{i \epsilon R_1} (y_i - \hat{y}_1)^2 + \sum_{i \epsilon R_2} (y_i - \hat{y}_2)^2$$

- The MSE or RSS of a tree with T regions is

$$RSS = \sum_{m=1}^{T} \sum_{i=1}^{r_m} (y_i - \hat{y}_m)^2$$

r_m : n. of observations in region m

CROSS VALIDATION (CV)

- The MSE or RSS of a tree with two regions R_1 and R_2 is

$$RSS = \sum_{i \epsilon R_1} (y_i - \hat{y}_1)^2 + \sum_{i \epsilon R_2} (y_i - \hat{y}_2)^2$$

- The MSE or RSS of a tree with T regions is

$$RSS = \sum_{m=1}^{T} \sum_{i=1}^{r_m} (y_i - \hat{y}_m)^2$$

r_m : n. of observations in region m

- CV minimizes

$$RSS = \sum_{m=1}^{T} \sum_{i=1}^{r_m} (y_i - \hat{y}_m)^2 + kT$$

k : shrinkage (complexity) parameter