Classification and Regression Trees

Lauren Flanakin
Outline

- Development of CART
- Definition
- CART Steps
- Visual Explanation
- Advantages/Disadvantages
- Examples
- Review
Development of CART

- Leo Breiman - as an Applied Statistician, he discovered tree-based methods of Classification that later became machine learning
- Wrote *CART: Classification and Regression Trees* with Jerome Friedman and Richard Olshen in 1984
Definition of CART

- CART: builds classification or regression trees for numeric attributes (regression) or categorical attributes (classification)
CART Steps

1. Start with root node (all data in dataset)
2. Split the node with max purity with “Gini”
   ○ Recursive process
3. Assign nodes with predicted classes
4. Missing data: program uses best available info to replace missing data (based on a variable that is relative to the outcome variable)
5. Stop tree building: when every aspect of the dataset is visible in decision tree
6. Tree Pruning: based on miscalculation rate
7. Optimal Selection: best tree that fits dataset with a low percentage of error
Visual Example

INPUT

Root node

Condition A

Node 1

Condition B

Node 3

Class

Node 4

Condition C

Node 5

Class

Node 6

Class

Node 2

Class
Advantages and Disadvantages

- **Advantages**
  - Handles data with any structure
  - Machine learning—little input from analyst
  - Final results can be summarized in logical if-then conditions

- **Disadvantages**
  - Knowing when to stop splitting
  - Computations are complex in determining best split conditions
Example of Classification Tree

Node 1
(Entire Group)
N = 150, W = 150
Species = Setosa
Misclassification = 66.67%

Node 2
Petal length ≤ 2.45
N = 50, W = 50
Species = Setosa
Misclassification = 0.00%

Node 3
Petal length > 2.45
N = 100, W = 100
Species = Versicolor
Misclassification = 50.00%

Node 4
Petal width ≤ 1.75
N = 54, W = 54
Species = Versicolor
Misclassification = 9.26%

Node 5
Petal width > 1.75
N = 46, W = 46
Species = Virginica
Misclassification = 2.17%
Example of Regression Tree

Node 1
(Entire Group)
N = 506, W = 506
Median value = 22.533
Std. dev. = 9.188

Node 2
Num. rooms ≤ 6.941
N = 430, W = 430
Median value = 19.934
Std. dev. = 6.346

Node 3
Num. rooms > 6.941
N = 76, W = 76
Median value = 37.238
Std. dev. = 8.929

Node 4
Lower status ≤ 14.4
N = 255, W = 255
Median value = 23.350
Std. dev. = 5.100

Node 5
Lower status > 14.4
N = 175, W = 175
Median value = 14.956
Std. dev. = 4.390

Node 6
Distance ≤ 1.35735
N = 5, W = 5
Median value = 45.580
Std. dev. = 8.840

Node 7
Distance > 1.35735
N = 250, W = 250
Median value = 22.905
Std. dev. = 3.858

Node 8
Crime rate ≤ 6.966832
N = 101, W = 101
Median value = 17.138
Std. dev. = 3.375

Node 9
Crime rate > 6.966832
N = 74, W = 74
Median value = 11.978
Std. dev. = 3.831
Review

- Development of CART
- Definition
- CART Steps
- Visual Explanation
- Advantages/Disadvantages
- Examples
- Review
References

- www.wikipedia.com