IIT ROORKEE

NPTEL

NPTEL ONLINE
CERTIFICATION COURSE

Pruning Process Part-2

LECTURE-43

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

CLASSIFICATION & REGRESSION TREES

- Pruning
  - Avoid overfitting
    - Full grown tree leads to complete overfitting of data
    - Poor performance on new data
  - Overall error of tree models
    - Expected to decrease until the point where relationships between outcome variable and predictors are fitted
    - Then tree models start fitting to the noise and overall error starts increasing
      - Due to splits involving small number of observations

CLASSIFICATION & REGRESSION TREES

- Pruning
  - Stop tree growth before it starts overfitting data or fitting noise
    - No. of splits or tree depth level
    - No. of observations in a node to attempt the split
    - Accepted level of reduction in impurity
    - Difficulties in determining the stopping point for such rules
  - Prune the full grown tree back to a level where it doesn’t overfit data or fit noise
    - Use validation partition to prune the tree modeled with training partition
    - Idea is to remove the tree branches which don’t reduce the error rate further

CLASSIFICATION & REGRESSION TREES

- Pruning
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Find the point where error rate on validation partition starts to increase
    - Cost complexity parameter or complexity parameter (CP) in CART algorithm

$$CP = Err + PF * TL$$

Where Err is misclassification error, PF is penalty factor for tree length (TL)

  - Minimum error tree
    - Tree with minimum misclassification error on validation partition

CLASSIFICATION & REGRESSION TREES

- Pruning
  - Best pruned tree
    - Adjustment for sampling error on minimum error tree
    - Smallest tree in the pruning sequence which lies within one std. err. (of error rate) of minimum error tree
- Open RStudio
- Classification Rules
  - Each terminal node in a tree model is equivalent to a classification rule
  - Simplify and remove redundant rules

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…