Machine Learning: Classification Tree

Nusrat Jahan, Department of Mathematics and Statistics

James Madison University, Harrisonburg, VA

Abundance of Data

- Advances in computational sector have allowed collection and processing of data of all kinds, in massive amounts
- Numerous processes are generating huge volumes of complex data
- Complexity is not only limited to the number of observations and variables, but also extends to individual cases in terms of information stored as curves, images, and videos
- Problems with high dimensional data analysis: composite design, multilevel curation, storage space problems, computational time & speed, etc.
- Traditional statistical analysis techniques are seldom applicable to high dimensional data because of violation of distributional and independence assumptions

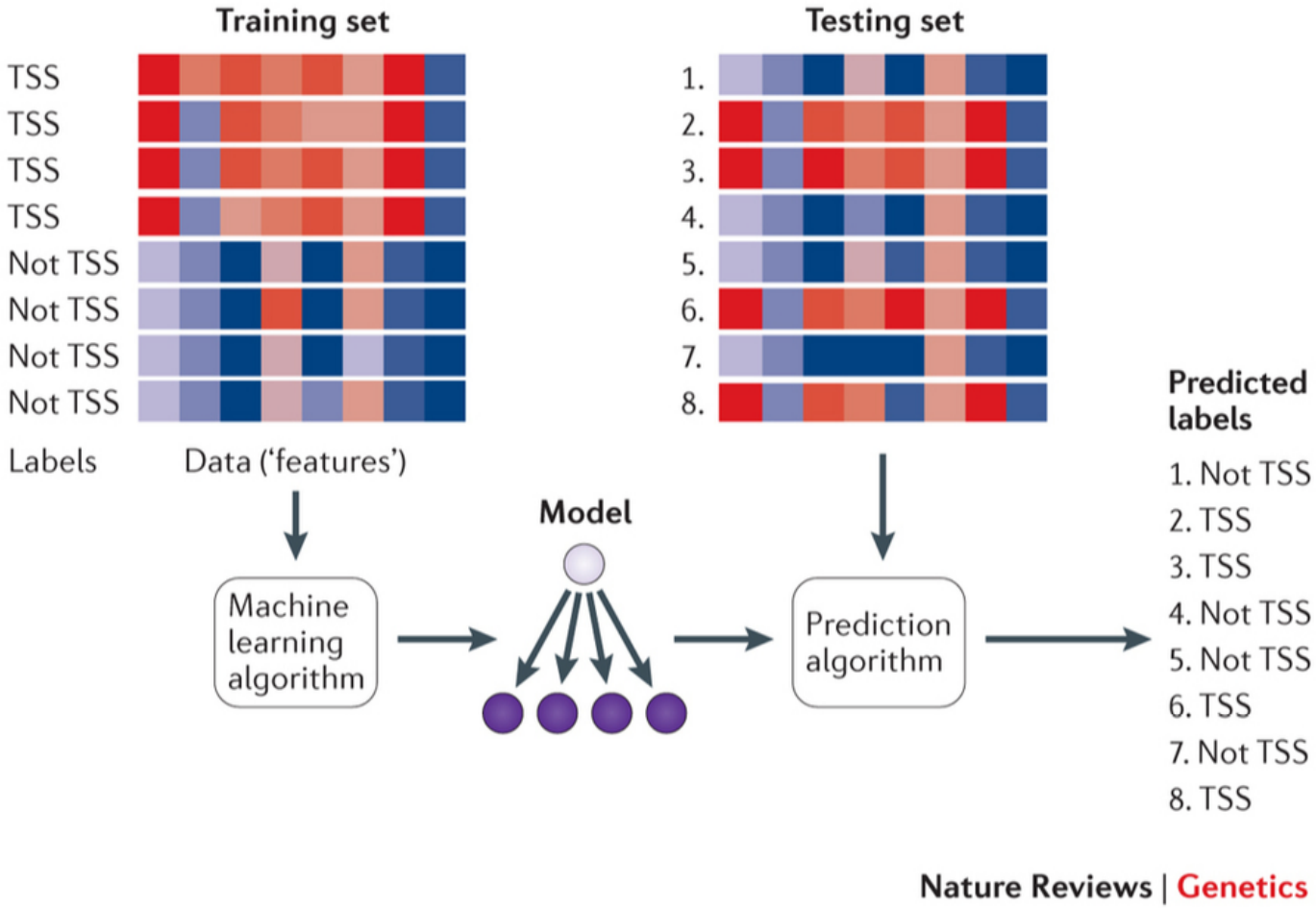
Learning From Data

- A data-driven decision making process
- Iterative process of class prediction for complex high dimensional data
- Initial prediction functions are developed based on subsets of data (training data)
- Final predictive function is an aggregate of all the initial functions
- Independently validated by exposing to a new data (test data)
- These approaches do not depend on the probabilistic distributional assumptions of data
- Very adept in handing data structures characterized by small sample size, high dimensional, and inter-dependencies

Machine Learning (ML):

- Data-driven iterative model selection process. Independent of distributional assumptions. Suitable for complex, large data analysis
- Methods are made possible with recent advances in computing power. Data is divided into two parts: training and test data sets, are part of the same population.
- Training set is used to develop a model and test set is used to evaluate that model performance.
- Models are used for prediction and inference.
- Model selection is done based on the bias vs. variance trade off.
- Two basic ML approaches: Supervised and Unsupervised Learning

Machine Learning Approach



Machine Learning Approaches

Supervised Learning

- Relates the response variable to the predictor variables. These relationships allow us to predict future values of the response variable, further investigate the functional relationship between the response and predictor variables. Examples: Classification tree, boosting, bagging, etc.

Unsupervised Learning

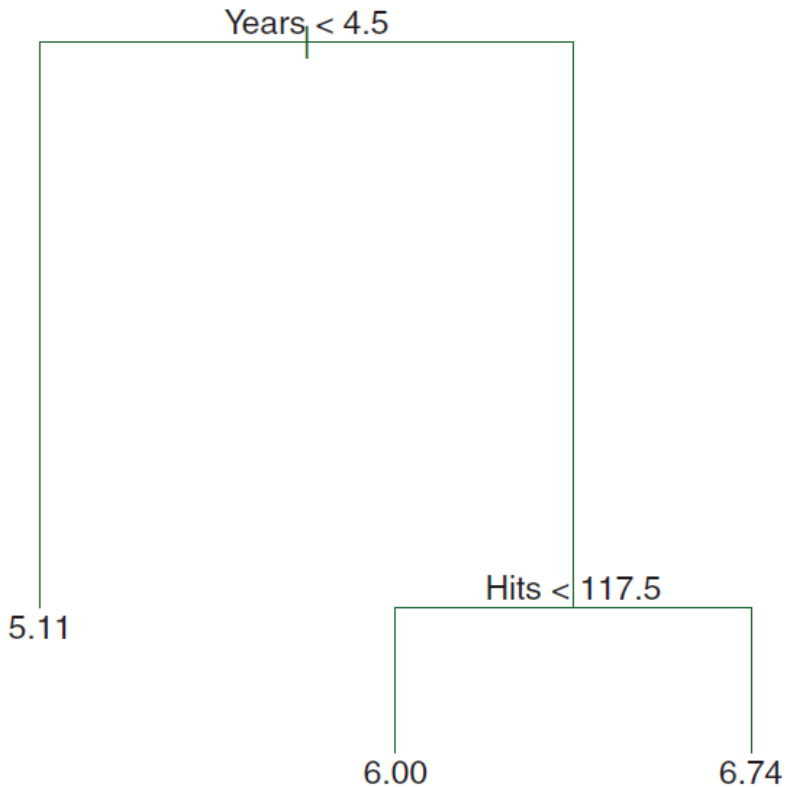
- These methods are used in situations when no clear response variable is available. All the variables are considered vectors of the measurements. Allows for the exploratory analysis of the relationships between the observations and variables. Examples: Cluster analysis, discriminant analysis, etc.

Tree based approaches

- An iterative process of building a classifier function sequentially from a subset of the data (training set)
- Classifier function $f(X)$ is constructed in such a way that, a binary spilt of the predictor space: $\{X|f(X) \leq s\}$ and $\{X|f(X) > s\}$ generates the largest reduction of prediction error
- Splitting process continues sequentially with a goal to minimize the prediction error
- Function parameters are updated at each iteration based on prediction errors
- Final classifier function is validated on a new data set (test data)

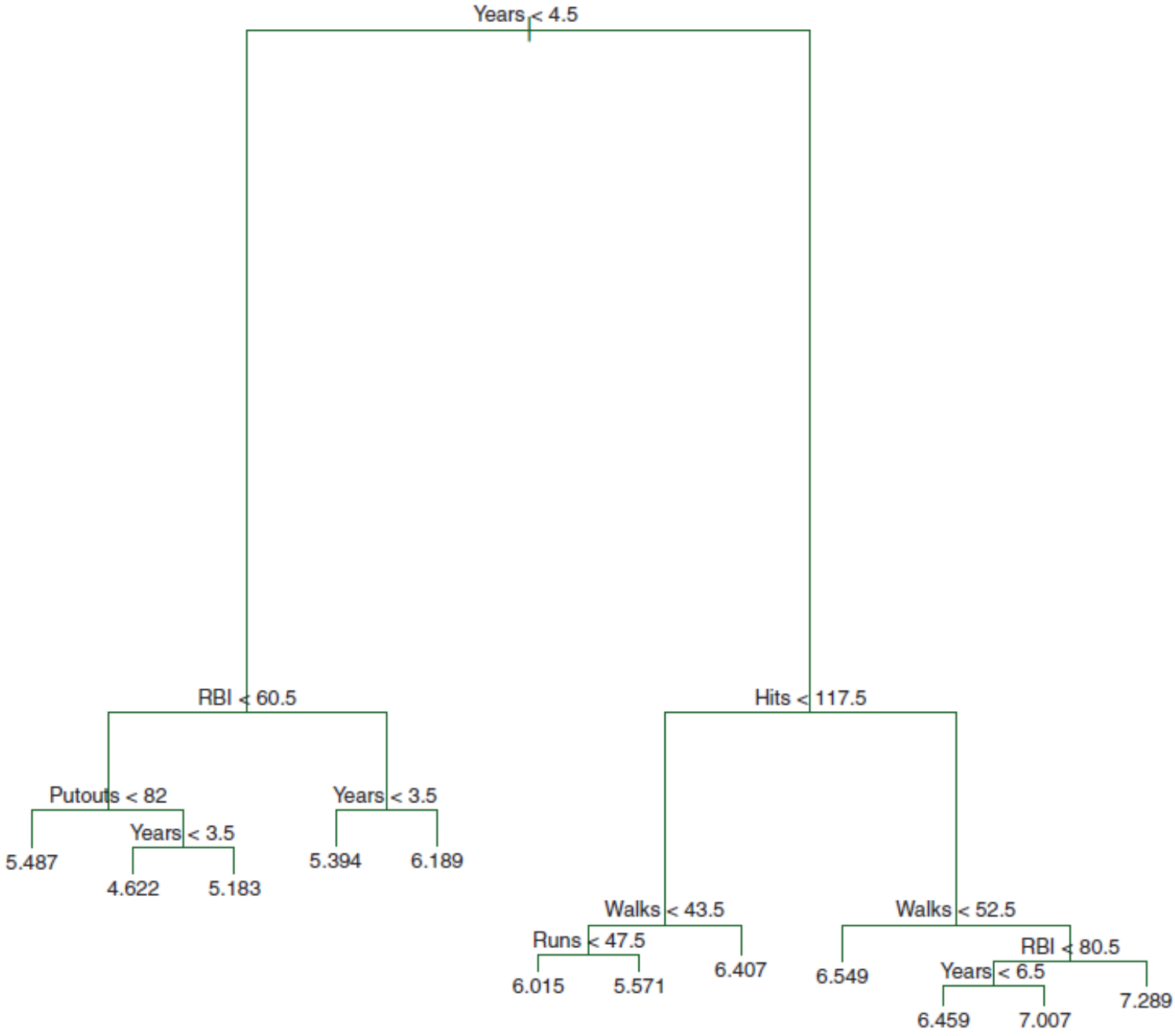
Example: Predict a baseball players Salary

- Number of years that he has played in the major leagues
- Number of hits that he made in the previous year



- Top split assigns Years < 4.5 to the left branch, predicted mean log salary is 5.107
- Players with Years $>= 4.5$ are assigned to the right branch, & further subdivided by Hits.

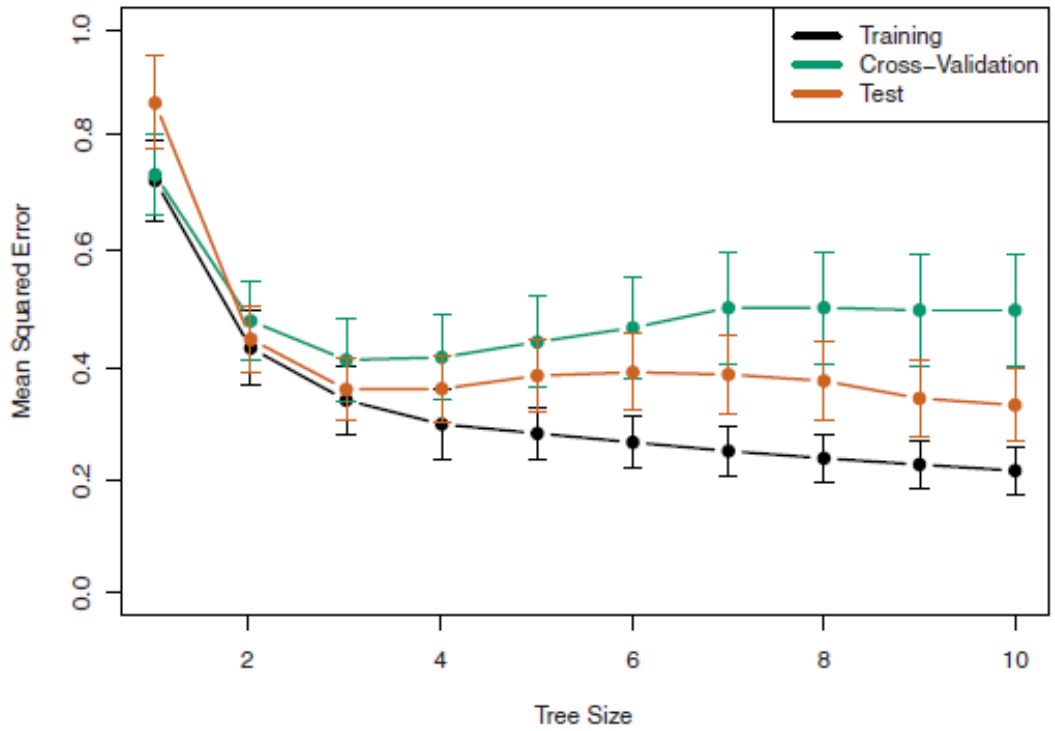
Classification Tree: Baseball Player's Salary



Cross-validation and Pruning

- CV is a resampling method to estimate model parameter values
- The training sample is divided into k folds
- For each $K = 1, 2, \ldots, k$, best subset of trees are obtained as a function of model parameter
- Mean square prediction error is evaluated in the left-out kth fold
- Estimated parameter values are chosen minimize the average error
- Based on the error rate, the tree can be pruned to include only a selected variables

Classification Tree: Baseball Player's Salary



- Tree has been pruned to 3 terminal nodes via cross-validation

Predict Sales of Car Seats

- "Carseats" is a built-in data set inside R
- It is a data set with 400 observations and 11 variables
- Data Source: rdrr.io/cran/ISLR/Carseats.html
- See the handout for detail information
- Reference: James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R

Useful Resources:

- R (Open source professional software for data analysis) https://cran.r-project.org/bin/windows/base/
- RStudio (Open source professional software for data analysis) https://www.rstudio.com
- https://www.statlearning.com, Springer-Verlag, New York