Multivariate Statistical Learning Using Random Forests

SDSU Bridges Summer 2018

Barbara A. Bailey

Department of Mathematics and Statistics
Computational Sciences Research Center
San Diego State University

Outline

- Statistical Learning
- Examples of Statistical Learning
- The Nonparametric Bootstrap
- Trees
- Random Forests
- Making Sense out of a Forest
- Iris Data

What is Statistical Learning?

- In artificial intelligence, machine learning involves some type of machine that modifies its behavior based on experience.
- In statistics, machine learning uses data to learn.
- Training data: (y, x)'s
  Two types: supervised and unsupervised learning

Some Examples of Statistical Learning

- Predict whether a patient hospitalized due to a heart attack will have second heart attack.
  Based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock 6 months in the future.
  Based on company performance measures and economic data.
- Identify numbers in handwritten ZIP codes.
  Based on digitized image.

Some Goals of the Statistical Analysis

- *Classification:* Group data based on predetermined classes, develop criteria for distinguishing between classes (Supervised Method)
- *Clustering:* Discover reasonable groupings within a dataset (Unsupervised Method)
- *Variable Selection:* Reduce the number variables required to perform a classification or clustering task, determine interrelationships between variables (can be Supervised or Unsupervised)

Example: South African Heart Disease Data

- 462 observations on males in South Africa
- Variable of interest is congestive heart disease where a 1 indicates the person has the disease, 0 he does not
- Explanatory variables include measurements on blood pressure, tobacco use, bad cholesterol, adiposity (fat %), family history of disease (absent or present), type A personality, obesity, alcohol usage, and age

- Question: How could you find the best predictors of heart disease?

Statistical Methods

- R
- Bootstrap
- Trees
- Random Forests

The Nonparametric Bootstrap

- What does nonparametric mean?
- What is bootstrapping and what is it good for?
  - Resampling technique used to obtain properties of estimators (summary statistics) from data
  - Uses random sampling with replacement

Trees

- What is a tree?
- Tree-based algorithms
- How to grow (and prune) a tree in R
- Example: South African Heart Disease Data

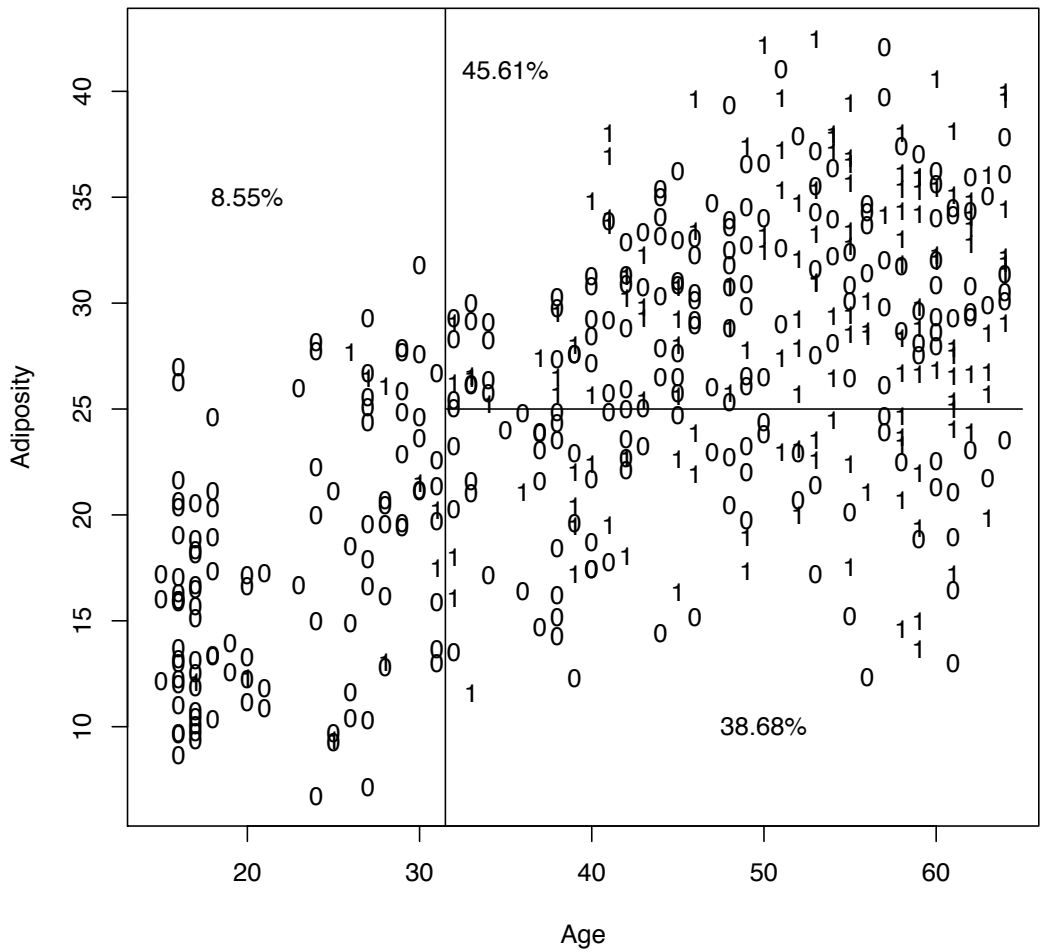


Figure 6.1: Splitting on age and adiposity

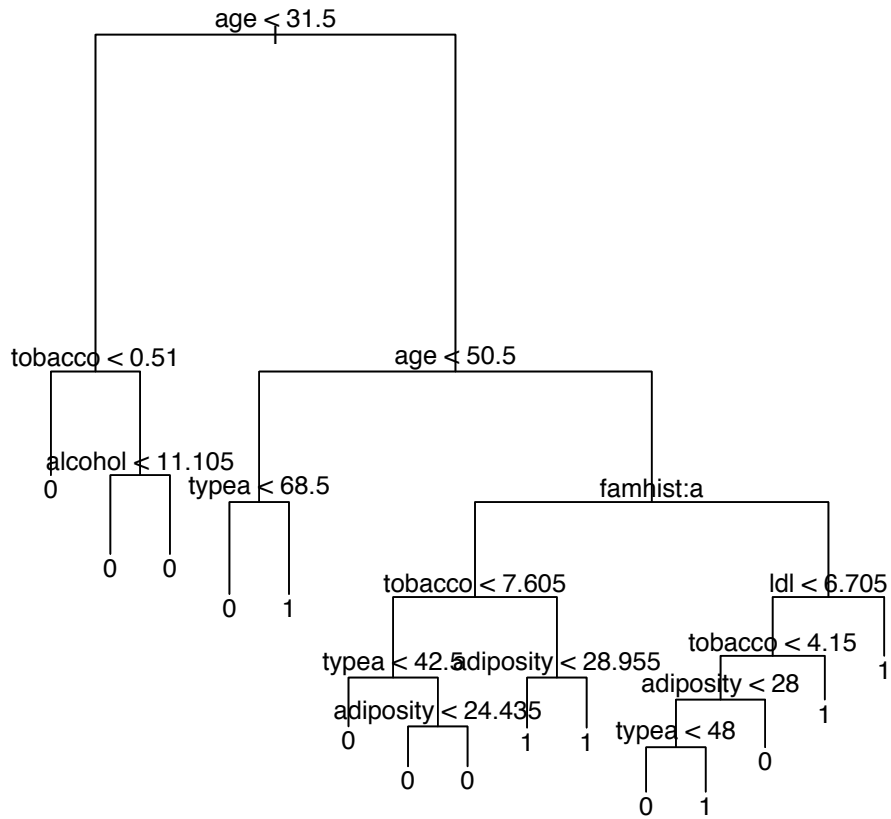


Figure 6.3: A large tree, with classifications at the leaves

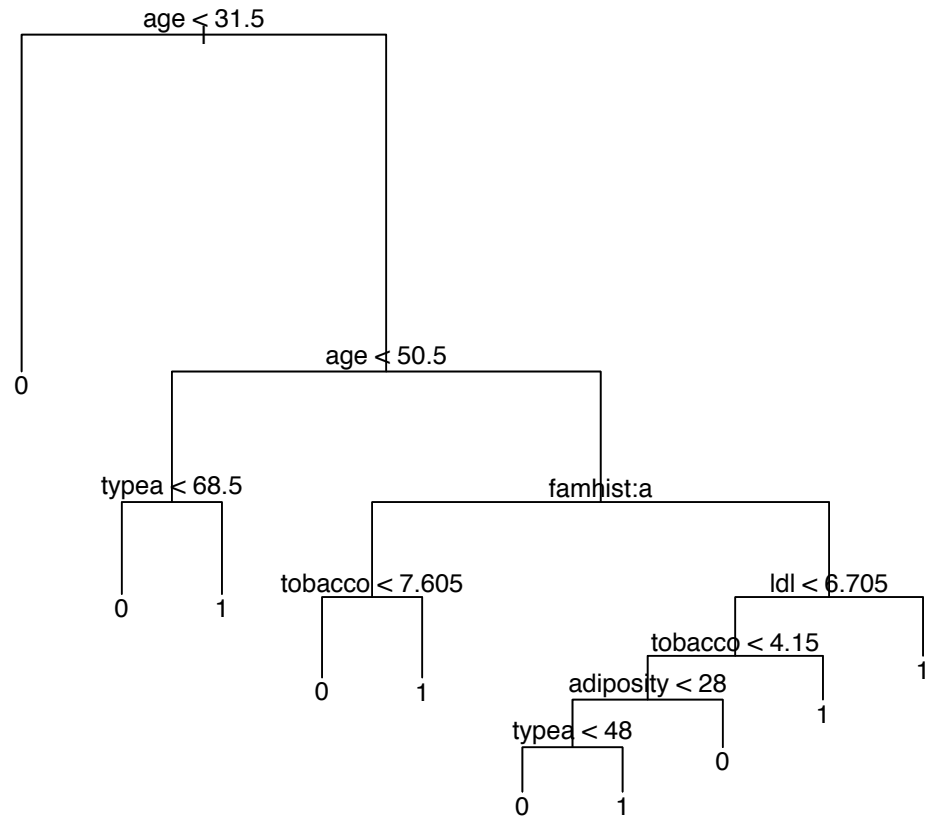
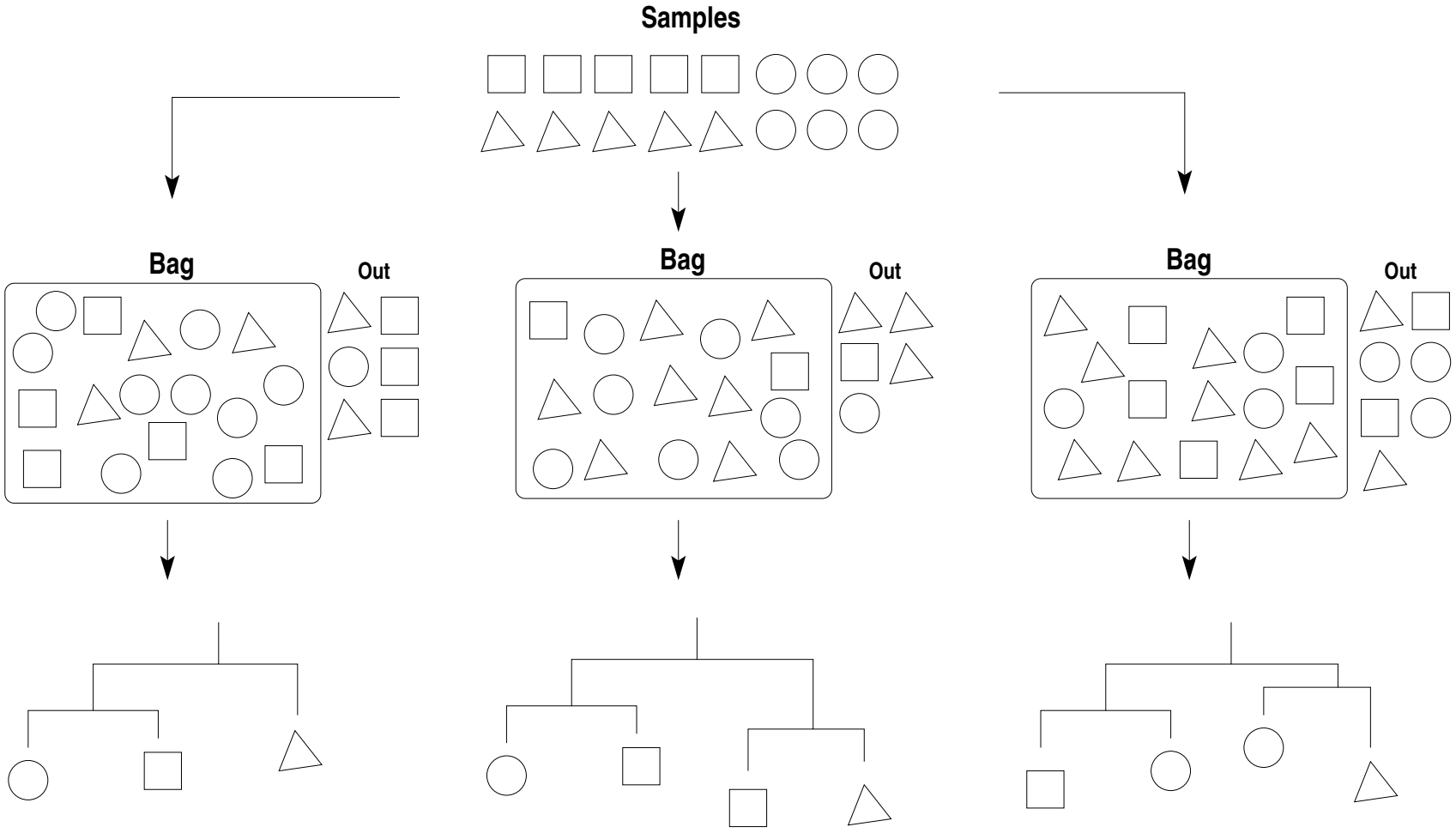


Figure 6.4: The tree, with unnecessary branches snipped

Random Forests

- A Random Forest is composed as a set of trees.
- Each tree in a Random Forest is generated from a random subset of all the data. This subset is generated by bagging: **b**ootstrap **ag**gregation - sampling with replacement. Unsampled data in each set are called *out-of-bag*.
- Each node in each tree is determined from a random subset of all the variables.
- Instead of classifying new data by tree branching rules, Random Forest classifies by vote of its component trees.

Random Forest Generation



Supervised and Unsupervised Random Forests

A Random Forest can be supervised or unsupervised.

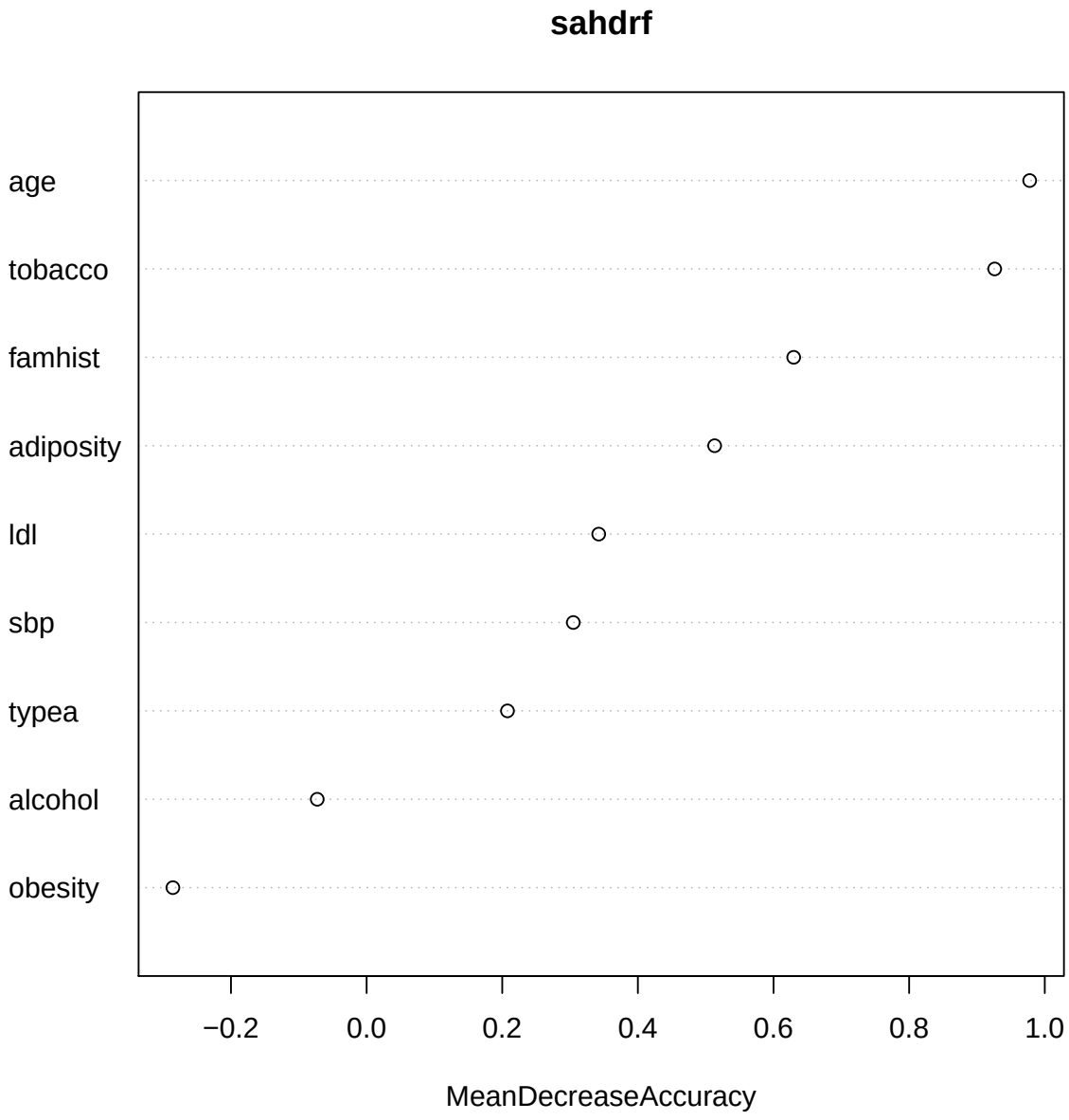
- Supervised:
  - In a supervised Random Forest, groupings for the training data are input to the algorithm.
  - Estimated classification error is computed using out-of-bag data.

RF: Variable Importance

Random Forests can report which variables were most important during construction. Particular variables are considered more important if:

- The accuracy of prediction of a sample is diminished when that particular variable in the sample is replaced with random noise during error analysis.
- The nodes of the trees become more homogeneous when that particular variable is used.

Variable Importance Plot



References for Trees (and more)

- Notes on Statistical Learning by John Marden