Introduction to Data Science and Statistical Learning Using Random Forests

CSRC Data Science Workshop
Summer 2019

Barbara A. Bailey

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

Outline

- What is Data Science?
- Statistical Learning
- The Nonparametric Bootstrap
- Trees
- Random Forests
- Making Sense out of a Forest

What is Data Science?

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.
- Machine Learning arose as a subfield of Artificial Intelligence
- Statistical Learning arose as a subfield of Statistics
- There is much overlap!
- 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.
- Classify pixels in a LANDSAT satellite image, by usage.

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)

Objectives of the Analysis

From training data:

- Accurately predict unseen test cases or data.
- Understand what and how inputs affect the outcome.
- Assess the quality of predictions and inferences.

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 and RStudio
- 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

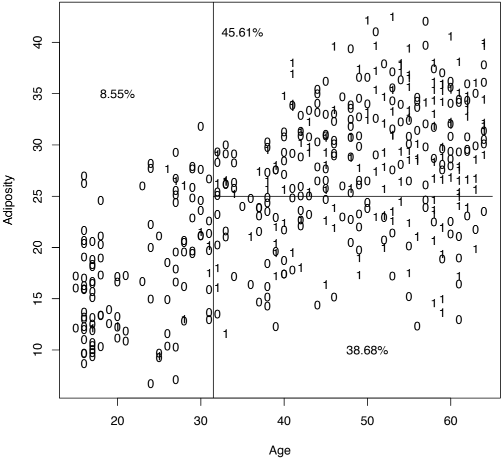


Figure 6.1: Splitting on age and adiposity

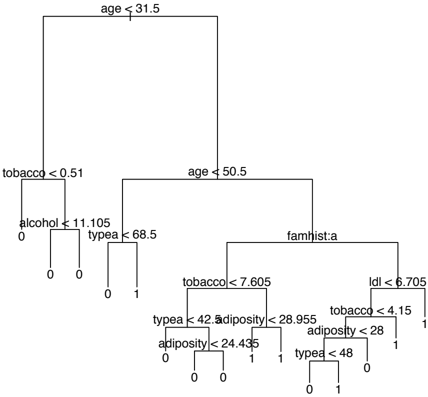


Figure 6.3: A large tree, with classifications at the leaves

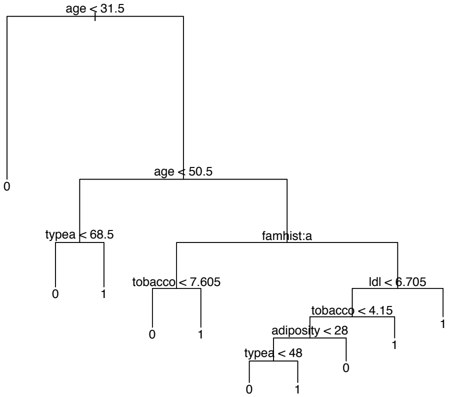


Figure 6.4: The tree, with unnecessary branches snipped

Random Forests

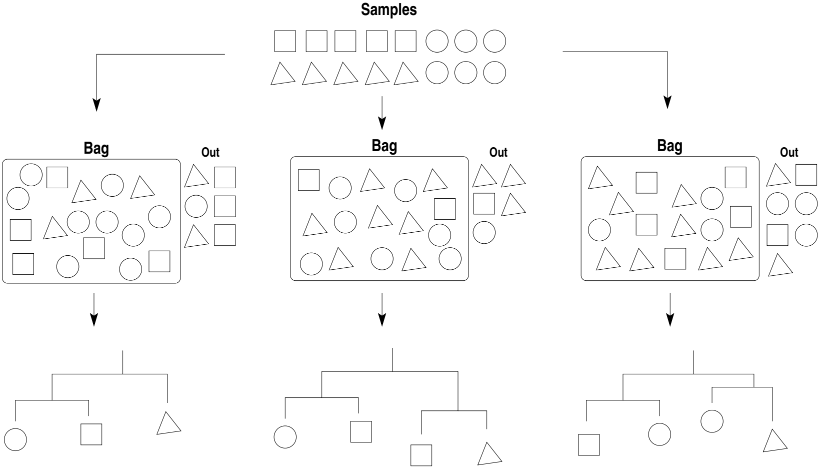
- A Random Forest is composed as a set of trees.

- 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.

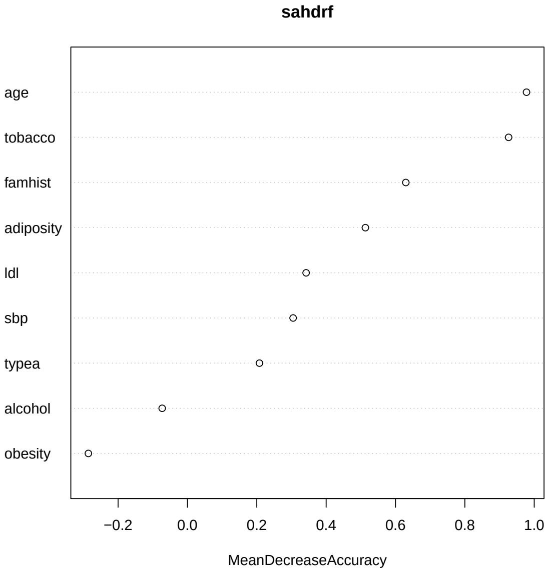
- Supervised:
  - In a supervised Random Forest, groupings for the training data are input to the algorithm.
  - Estimated classification error is computed using out-of-bag data.

RF: Variable Importance

Random Forest reports which variables are 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



Some References

- An Introduction to Statistical Learning (ISLR) at www.StatLearning.com, by James, Witten, Hastie, and Tibshirani
- The Elements of Statistical Learning by Hastie, Tibshirani, and Friedman (more advanced)
- Notes on Statistical Learning by John Marden (even more advanced)

Part II

- Random Forest for Regression Estimating
- Model Selection
- Model Assessment

The Model

Suppose we have response Y and p different predictors $X = (X_1, X_2, \ldots, X_p)$.

We can write the model: $Y = f(X) + e$

where e is random noise or an error term, which is independent of X and has mean zero.

The regression function is:

$$f(x) = f(x_1, x_2, \ldots, x_p) = E(Y | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

How do we estimate f? ($\hat{f}$ is our estimate for f)

Model Selection and Model Assessment

- *Model Selection:* Estimating performance of "different" models in order to choose the "best" one.
- *Model Assessment:* Having chosen a final model, estimate its prediction error on new data. (Generalization Error)

Next set of slides is from the ISLR book!

Assessing Model Accuracy

Suppose we fit a model $\hat{f}(x)$ to some training data $\mathsf{Tr} = \{x_i, y_i\}_1^N$, and we wish to see how well it performs.

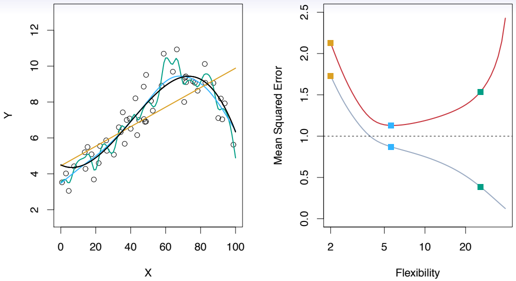
- We could compute the average squared prediction error over Tr:

$$\text{MSE}_{\mathsf{Tr}} = \text{Ave}_{i \in \mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$$

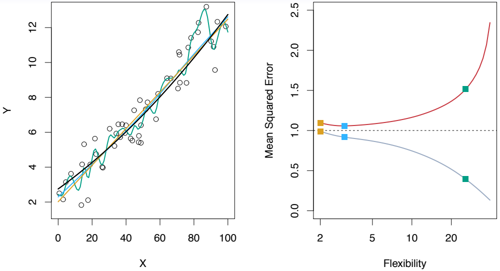
This may be biased toward more overfit models.

- Instead we should, if possible, compute it using fresh *test* data $\mathsf{Te} = \{x_i, y_i\}_1^M$:

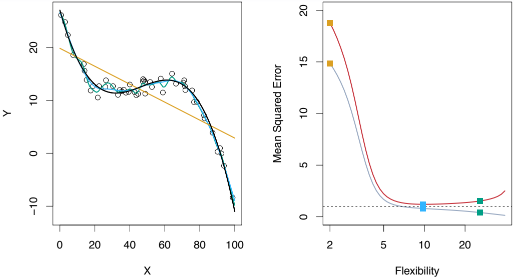
$$\text{MSE}_{\mathsf{Te}} = \text{Ave}_{i \in \mathsf{Te}}[y_i - \hat{f}(x_i)]^2$$



Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.



Here the truth is smoother, so the smoother fit and linear model do really well.



Here the truth is wiggly and the noise is low, so the more flexible fits do the best.

Bias-Variance Trade-off

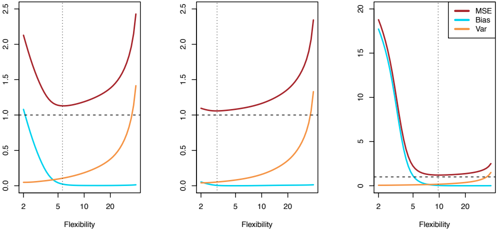
Suppose we have fit a model $\hat{f}(x)$ to some training data Tr, and let (x_0, y_0) be a test observation drawn from the population. If the true model is $Y = f(X) + \epsilon$ (with $f(x) = E(Y|X = x)$), then

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

The expectation averages over the variability of y_0 as well as the variability in Tr. Note that $\mathrm{Bias}(\hat{f}(x_0))] = E[\hat{f}(x_0)] - f(x_0)$.

Typically as the *flexibility* of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a *bias-variance trade-off.*

Bias-variance trade-off for the three examples



Part III

- Unsupervised Learning - Discover interesting things about the measurements or features.
  - *PCA:* Principal Component Analysis for Dimension Reduction (not covered here)
  - *Clustering:* Discover reasonable groupings within a dataset

Unsupervised Random Forests

An unsupervised RF can be used to estimate a proximity matrix for clustering.

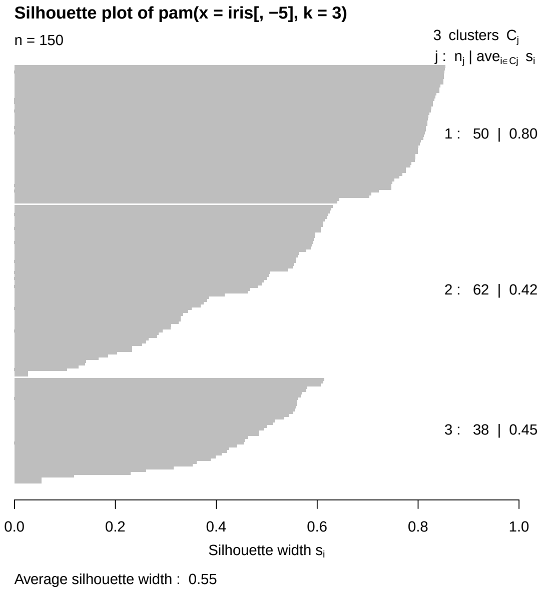
- The (i, j) element of the matrix is the fraction of trees that i and j fall in the same terminal node.
- Trick:
  - Call original data "class 1".
  - Generate synthetic "class 2" data by sampling uniformly within the range of each variables.
  - Use supervised RF on the above 2 classes to estimate the proximity matrix.

Clustering with the Proximity Matrix

- We choose Partioning around Medoids (PAM)
  - Similar to k-means but uses the median.
  - More robust to outliers and noise.
  - Choose the "best" number of classes using silhouettes.

Silhouettes

- Can be used with any clustering algorithm.
- Description for each proposed clusters number k:
  - For each data point, first find the average distance between it and all other points in the same cluster.
  - Then find the average distance between the data point and all points in the nearest cluster.
  - The silhouette coefficient for each data point is defined as the difference between the above, divided by the greater of the two.
  - Use the average silhouette coefficient to obtain an "overall" measure.
- Calculates a measure of dissimilarity (so high is good).
- Use average silhouette plot over a range of the number of clusters k to determine best number of groups.



Silhouette plot of pam(x = iris[, −5], k = 3)
n = 150
3 clusters C_j
j : n_j | $\text{ave}_{i \in Cj}$ s_i
1 : 50 | 0.80
2 : 62 | 0.42
3 : 38 | 0.45
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.55