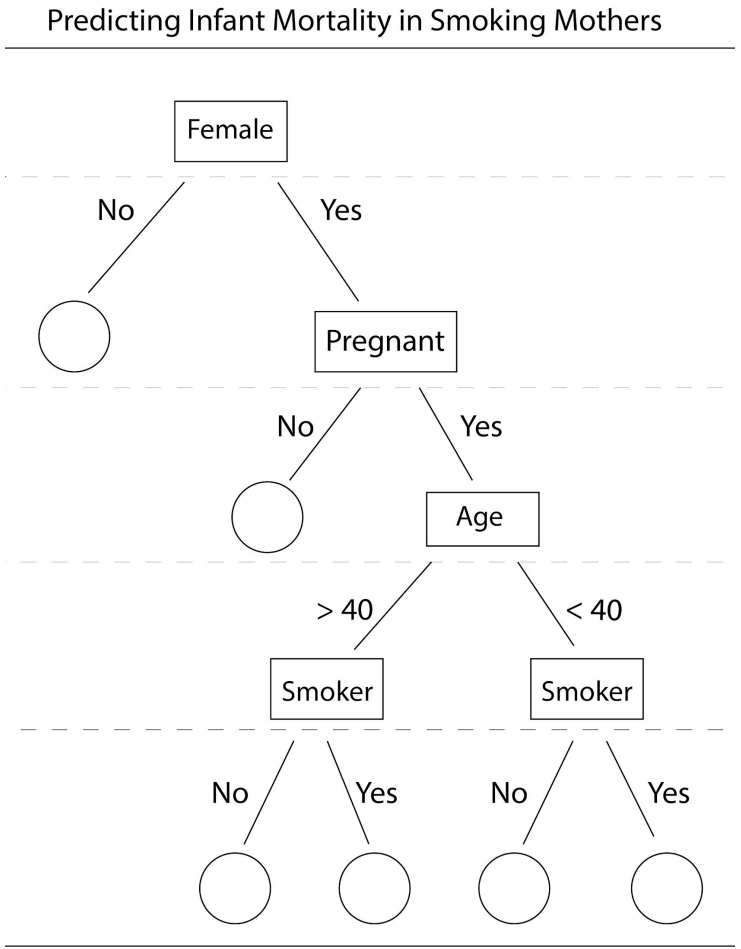
Decision Trees & Random Forests

Samir Rachid Zaim
‹samirrachidzaim@email.arizona.edu›

What is a decision tree?

This

Not this



What is a decision tree? Decision Trees 101.

A decision tree can be thought of as a series of yes or no questions. At each step of the decision making process, we split our sample by how they answer each question.

Trees can be equally applied to regression as they can to classification.

Every data point starts at the root node, and ends at a terminal (leaf) node.

Every node that is not a terminal node is a splitting node.

How do we form good questions?

Good Questions

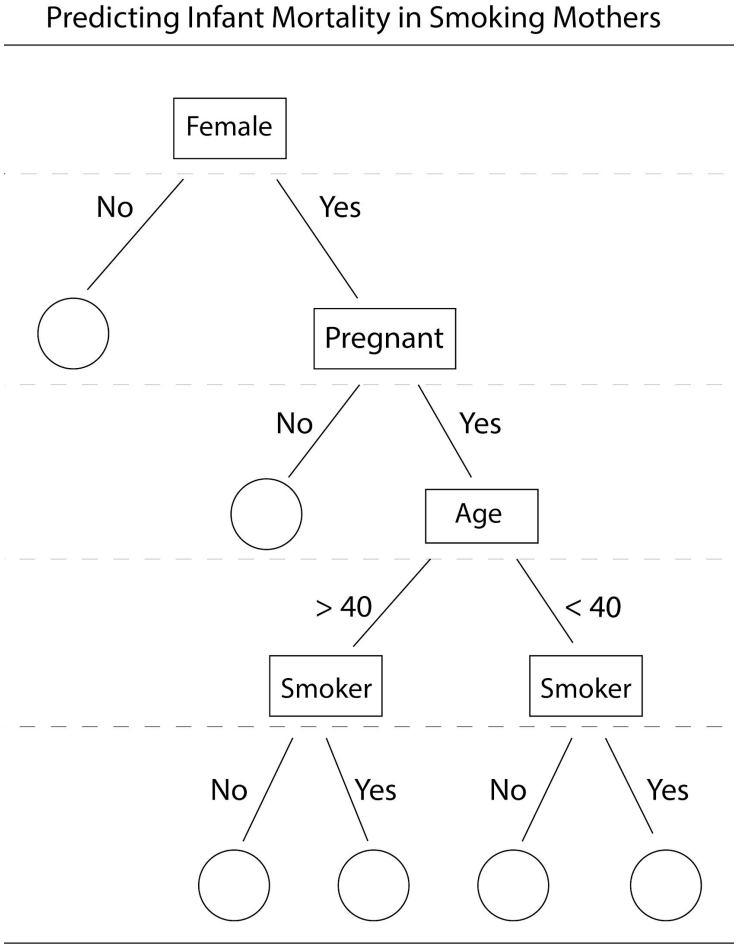
- Is this person on medicare?
- Does this person have septic shock?
- Is this basketball player a good 3pt shooter?
- Did this team win their conference?

Bad Questions

- How are you feeling today?
- How did you feel about the movie Black Panther?

How do we go from a question to a tree?

How do we choose which question goes at which splitting node?



Classification v. Regression Trees

Regression trees ≈ Linear Regression

- Predict mean response for a data point(i.e. stock market value)
- Each split is evaluated using a continuous loss function (i.e. Least Squares)
- The mean response prediction is the average of the observations in the terminal node
- Code example: tree
  - tr = tree(y ~., cbind(X,y))
  - plot(tr)
- Code example
  - rf1 = randomForest(Xtrain, y)
  - rf.preds = predict(rf1 , Xtest, 'prob')
  - important.vars = importance(rf1)

Classification v. Regression Trees (2)

Classification trees ≈ Logistic Regression

- Predict class label for each data point (i.e. ice cream flavor preference)
- Each split is evaluated using a discrete loss function (entropy, 0-1 loss function, weighted loss function, etc..)
- The class label prediction is done by majority vote
- Code example: tree
  - tr = tree(Species ~., iris)
  - plot(tr)
- Code example: Random Forest
  - rf1 = randomForest(Xtrain, factor(y))
  - rf.preds = predict(rf1 , Xtest, 'prob')
  - important.vars = importance(rf1)

Should we trust a single tree?

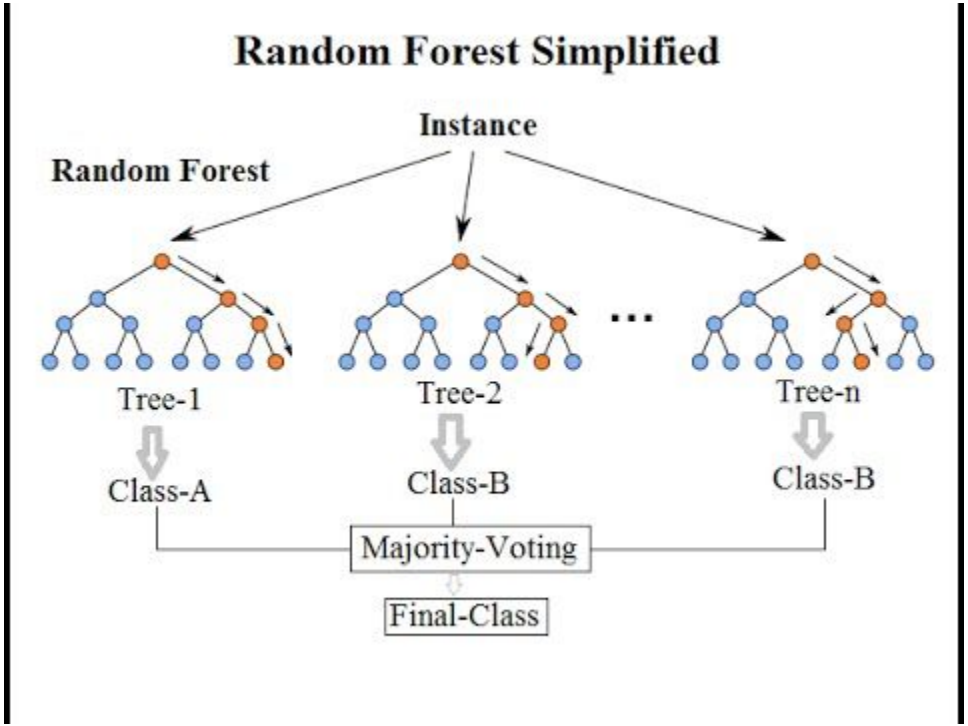
What are the pros and cons of a single model?

What are the pros and cons of an ensemble model?

How does variable selection work in each?

Can we get an accurate representation of performance?

Extending Decision Trees to Random Forests



Extending Decision Trees to Random Forests (2)

To get randomized and diverse predictions for each tree:

- Bootstrap a % of observations for each tree at random
- At each splitting node, subset a % of the features/predictors

Then aggregate the random forest to get a consensus prediction for each observation:

- Regression → take mean prediction of all trees
- Classification → Majority vote amongst all tree predictions

Hands on Activity

Using Random Forests to Classify Playoff Success

Goal: Predict Playoff Appearance

- Part 1:
  - Develop your own decision tree by hand
  - What questions would you ask? Does the order in which the questions you ask matter?
- Part 2:
  - Using the tree package in R:
    - Create and plot a classification tree predicting playoff appearance
  - Using the randomForest package in R:
    - determine each team's probability of making playoffs using 5, 10, 20,50, 100 trees, using the entire dataset for training and testing (Bad Practice! In general but just as an example)
    - Identify the important variables in each model and compare them as the forest grows deeper trees