Trees: stratify the prediction space in order to make predictions

* Called decision trees
* Simple, easy to interpret
* Accuracy is not as good as other methods
  + Trees are an oversimplification of the model
  + We can increase accuracy with bagging, boosting, random forest
* Trees work for regression and classification
* Terminal nodes: regions that are partitioned within the data
  + Terminal nodes are the final ones that appear at the bottom of the tree
* Internal nodes: points in the tree where the predictor space is split
* In drawn trees, leaves are at the bottom
* First split is done on the most important variable
  + Hitters dataset example: Years is the most important
* Regression trees: prediction in each region is the average of the response values in that region
  + Goal is still to minimize the RSS
* Tree algorithm takes a top-down, greedy approach
  + Can’t feasibly examine every possible combination of boxes
  + Top-down: start at the top with the whole set of observations, make 1 split
  + Greedy: take the split that gives the greatest improvement right now

Pruning trees: reduce the size of the tree to reduce overfitting

* Bushy tree: high variance, low bias, overfitting, not a good predictor
* Pruning strategy: grow a very large tree (until no node has more than perhaps 5 observations), then reduce the number of nodes
* Set an objective function:
  + T: number of terminal nodes in the tree
  + Use cross-validation to find the best value of alpha (the model with the smallest MSE)
    - Large alpha: large penalty for big trees, will result in a small tree
    - Small alpha: small penalty, tree can be bushy
* In the drawn tree, long tree arms mean the split resulted in a large decrease in RSS

Classification trees: used to predict categorical variables

* Predict that each observation belongs to the most commonly observed class in the region
* Can’t use RSS to assess classification trees
  + Error rate: fraction of training observations in the region that don’t belong to the most common class
    - Not sensitive enough to result in good trees
  + Gini index: measure of variance across the classes
    - Each term is a binomial variance
    - Takes on a small value if all observations fall in the same class
    - Maximum value when there is equal distribution among multiple classes
    - Also called purity index
  + Deviance/cross entropy: similar to Gini index

Bagging (bootstrap aggregation): method for reducing variance in a statistical model

* Idea: taking the average of a set of observations reduces variance
* Bagging: take bootstrap samples, grow a tree on each one, take the average of all of them
  + Regression trees: average the prediction for each region of the tree
  + Classification: take a majority vote to predict each point
* In bagging, no need to worry about pruning
  + Taking the average reduces variance, so pruning isn’t necessary
* Out of bag error estimate: analogous to LOOCV
  + Each bagged tree uses an average of 2/3 of the observations
  + Can use the remaining 1/3 of the observations as a test set
  + Results in B/3 predictions for each data point (B is the number of trees)

Random forests: similar to bagging, but adds a step to decorrelate the trees

* Decorrelation reduces variance when the trees are averaged
* Do bagging as normal, but for each split, only consider a subset of predictors to do the split
  + M: number of predictors to consider
  + Usually m = sqrt(p)
  + Take a fresh selection of m predictors at each split
* Random forest forces the trees to consider all predictors

Boosting: sequential algorithm to improve tree performance

* Each tree is fit to the previous tree’s residuals
* Procedure:
  + Fit a tree with d splits to the training data (X, r)
    - X: data
    - R: current residuals
  + Update the tree by adding a shrunken version of the current tree
  + Update the residuals
* Idea: learns slowly, which therefore prevents overfitting
* By fitting to residuals, improve the tree in places where it isn’t doing well
* Tuning parameters for boosting:
  + Number of trees B
  + Shrinkage parameter lambda (learning rate)
  + Number of splits d
    - If d=1, it’s a set of stumps
    - Might try 1, 2, 4, 8
    - D splits can include at most d variables
* State-of-the-art technique but can be hard to interpret

Variable importance: measure of determining how important each variable is for making predictions

* Regression trees: record total amount the RSS decreases after splitting on a given predictor, averaged over all B trees
* Classification trees: total sum of decreases in the Gini index, averaged over all B trees