IDS 702: Module 8.3

Ensemble tree methods

Dr. Olanrewaju Michael Akande

Bagging

- Instead of using one big tree, bagging constructs B classification and regression trees using B bootstrapped datasets.
- Each tree is grown deep and has high variance, but low bias.
- Averaging all B trees reduces the variance.
- Improve accuracy by combining hundreds or even thousands of trees.
- To predict,
  - a continuous outcome, drop new X down each tree until getting to terminal leaf. Predicted value of Y is the average of all B predictions across all the trees.
  - a categorical outcome, select the most commonly occurring majority level among the B predictions.

Random forests

- The trees in bagging would be correlated since they are all based on the same data (sort of!).
- Random forests attemps to de-correlate the trees.
- Random forests also constructs B classification and regression trees using B bootstrapped datasets but only uses a sample of the predictors for each tree.
- Doing so prevents the same variables from dominating the splitting process across all trees.
- Both bagging and random forests will not overfit for large B.

Random forests

- Random forest algorithm:

  For $b = 1, \ldots, B$,

  1. Take a bootstrap sample of the original data.
     - Alternatively, can take a sub-sample of the original data of size $m < n$, where n is the sample size of the collected data.
  2. Take a sample of $q < p$ predictors, where p is the total number of predictors in the dataset.
  3. Using only the data in the bootstrapped sample or sub-sample, grow a tree using only the q sampled predictors. Save the tree.

- For predictions, do the same thing as in bagging.
- Variable importance measures based on how often a variable is used in splits of the trees.

Random forests vs. parametric regression: benefits

- No parametric assumptions.
- Automatic model selection.
- Multi-collinearity not problematic.
- Can handle big data files, since trees are small.
- In R, use the `randomForest` package.

Random forests vs. parametric regression: limitations

- Regression prediction limitations like those for CART.
- Hard to assess chance error.
- Little control over the few parameters to tweak if model does not fit the data well.

Boosting

- Boosting works like bagging, except that the trees are grown sequentially.
- Specifically, each tree is grown using information from previously grown trees.
- After the first tree, the remaining trees are built using residuals as outcomes.
- The idea is so that boosting can slowly improve the model in areas where it does not perform well.
- Boosting does not involve bootstrap since each tree is fit on a modified version of the original data set.
- It can overfit if the number of trees is too large.
- There are so many boosting methods! This is just one of them.

Boosting

- Goal: to construct a function $\hat{f}(y|x)$ to estimate true $f(y|x)$.
- Boosting algorithm:
  1. Fit a decision tree $\hat{f}$ with d splits to the data using Y as the outcome. Compute the residuals.
  2. For $b = 2, \ldots, B$,
     - Fit a decision tree $\hat{f}^b$ with d splits to the data using the residuals as the outcome.
     - Add this new decision tree into the fitted function: $\hat{f} = \hat{f} + \lambda \hat{f}^b$.
     - Compute updated residuals.
  3. Output the boosted model: $\hat{f} = \sum_{b=1}^{B} \lambda \hat{f}^b$.
- The shrinkage parameter λ (often small, e.g. 0.01) controls the rate at which boosting learns.

General advice about tree methods vs parametric regressions

- When the goal is prediction and sample sizes are large, tree methods can be effective engines for prediction.
- When the goal is interpretation of predictors, or when sample sizes are modest, use parametric models with careful model diagnostics.
- Either way, always remember the data:
  - What population, if any, are they representative of?
  - Are the definitions of variables what you wanted?
  - Are there missing values or data errors to correct?

Arsenic example again

- Recall the study measuring the concentrations of arsenic in wells in Bangladesh.
- We already fit a logistic regression to the data.
- We will use the same data to compare these models.
- Research question: predicting why people switch from unsafe wells to safe wells.
- The data is in the file `arsenic.csv` on Sakai.

Arsenic example again

Variable	Description
Switch	1 = if respondent switched to a safe well 0 = if still using own unsafe well
Arsenic	amount of arsenic in well at respondent's home (100s of micro-grams per liter)
Dist	distance in meters to the nearest known safe well
Assoc	1 = if any members of household are active in community organizations 0 = otherwise
Educ	years of schooling of the head of household

- Treat switch as the response variable and others as predictors.
- Move to the R script here.

What's next?

Well.....nothing!

You made it to the end of this course.

Hope you enjoyed the course and that you have learned a lot.