Class 4 Review Notes

AI & Machine Learning
Fall 2023 - Laurie Ye

4 More About Trees : Ensemble Methods, Bagging and Random Forests

4.1 Ensemble

The basic idea of ensemble methods in supervised machine learning is to create multiple prediction functions:

$$\hat{f}_1(\mathbf{X}), \hat{f}_2(\mathbf{X}), \ldots, \hat{f}_m(\mathbf{X})$$

These prediction functions are then "combined" to create an overall prediction function, usually in a simple way. The average of the prediction functions is often used:

$$\hat{f}^*(\mathbf{X}) = \frac{1}{m} \sum_{i=1}^{m} \hat{f}_i(\mathbf{X})$$

Medians, weighted averages or other measures of location could be used as well. In classification, multiple classifiers can be used to "vote" on the classification. In ensemble methods, the multiple prediction functions come about in four primary ways:

1. Use of different statistical methods (e.g., use of linear regression, regression trees, k-NN, etc.)
2. Use of the same basic statistical method, but with different sets of x-variables.
3. Use of the same basic statistical model but with different values of a tuning parameter (e.g., cost-complexity coefficient in regression trees, tuning parameter in regularization).
4. Use of different training data sets. For example, randomly drawing different training data sets.

Note 1. In ensemble, statistical models built sequentially with each model in the sequence trying to improve upon the error of the previous models. This approach is used in "boosting."

4.2 Bootstrap

- The bootstrap was created as a way of obtaining non-parametric estimates of the standard errors of model parameters (such as regression coefficients). - estimate how much error there might be in the calculations of certain statistics (like regression coefficients) without making strict assumptions about the data.

- Basic idea: randomly draw samples with replacement from a data set that are the same size as the data set.
- In the case of estimating the standard error of a model parameter, the parameter is estimated for each of the randomly drawn "bootstrap" samples. - When we want to find out how much our estimate of a model parameter (like a regression coefficient) might be off by, we use the bootstrap method. This involves making lots of new sample groups from our original data by randomly picking data points (with the possibility of picking the same one more than once) and then calculating the model parameter for each new group. By doing this many times, we can get a good idea of how much our parameter estimate might vary, which tells us about its standard error.
- The sample standard deviation of these estimates is then used to estimate the parameter's standard error.
- The theoretical idea behind bootstrap estimates of parameter standard error is based on the idea that the empirical distribution defined by the sample should be close to the true underlying distribution. - In that case, the bootstrap estimate of the standard error should be "close" to the true standard error.
- This is basically a "continuity" of distributions idea.

  Bootstrapping is a very useful and important approach.

  Note: Each bootstrap sample will use about $\frac{2}{3}$rds of the data.

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = \frac{1}{e} \approx 0.368$$

  as the number n becomes very large, the expression $\left(1-\frac{1}{n}\right)^n$ approaches the value $\frac{1}{e}$, which is approximately 0.368. This concept is related to the bootstrap method in statistics. When creating bootstrap samples, on average, each sample will contain about two-thirds of the unique data points from the original dataset, due to this statistical property. This is a statistical property that arises because as the number of samples n gets very large, the probability of any single data point being selected at least once in a bootstrap sample tends toward $1-\frac{1}{e}$, or approximately 63.2%.

4.3 Bagging

- "Bagging" stands for "bootstrap aggregation."
- In bagging, multiple prediction functions are created by applying a model building procedure (like regression or classification trees) to multiple bootstrap samples.
- The prediction function for each bootstrap sample will be different.
- These prediction function are then averaged.
- Bagging tends to "smooth out" or 'average out" both discontinuities in the model (e.g., the plateaus in regression trees) as well as overfitting.
- Bagging is a very important method for regression and classification.

4.4 Random Forests

- The regression trees built for different bootstrap samples in the "bagging" approach often follow essentially the same sequence of variable splitting.
- That is, virtually all (or even all) of the bootstrap samples result in the same variables split first, second, third, and so on.
- As a result, the "path" through the variables taken in all of the bootstrap samples is either identical or very similar.
- The idea of random forests is to improve the overall prediction power of bagging by "forcing" different paths through the variables.
- Forcing different paths is done as follows:
  1. A bootstrap sample is randomly selected.
  2. A subset of the x-variables is randomly selected. The number of variables selected is generally $\frac{p}{3}$ for regression or $\sqrt{p}$ for classification.
  3. The first split is determined using only the randomly selected x-variables. So only one of these variables can be used for the first split.
  4. Another sample of $\frac{p}{3}$ or $\sqrt{p}$ variables is selected. The second split is determined using only these variables.
  5. The tree is built continuing this process. Each split is restricted to considering only the x-variables randomly drawn at that step.
  6. When the tree is completed for a bootstrap sample, a new bootstrap sample is drawn and the procedure is repeated.
  7. Once the desired number of trees have been created, predictions are made by averaging the predictions from each of the individual trees.

Note 2. Imagine you're trying to guess the number of candies in a jar. You could just guess on your own, or you could ask a bunch of friends to guess too and then use everyone's guesses to make a better estimate. That's kind of what a random forest does with decision trees to make predictions or decisions.

Now, let's add in bagging and bootstrapping:

Bootstrapping (Making the Teams): Before your friends guess, you give each one a small bag of candies from the jar to look at. They use just what's in their bag to make their guess. This is bootstrapping: creating many small, random samples from your big jar (the original data) to help each friend (each decision tree in the forest) make a better-informed guess.

Bagging (Combining Guesses): This is like taking the average of all your friends' guesses about the candies. It's short for "bootstrap aggregating". By averaging, you're probably going to get closer to the real number than just one person's guess.

So, a random forest uses bootstrapping to create lots of different groups of data for the trees to learn from, and then it uses bagging to combine all the different trees' decisions into one final decision. This way, the random forest can make a really good estimate or decision because it's using the wisdom of a whole crowd of trees, rather than just one.

4.5 Reference

Goizueta Business School-Emory University: Professor George S. Easton