Chapter 1

Bagging and Random Forests

Bagging (Bootstrap Aggregating) and Random Forests are ensemble machine learning methods that are primarily used to improve the stability and accuracy of prediction models.

1.1 Bagging

Bagging, an abbreviation for Bootstrap Aggregating, is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages the output (for regression) or performs a vote (for classification).

Given a standard training set D of size n, bagging generates m new training sets D_i, each of size n', by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If $n' = n$, then for large n the set D_i is expected to have the fraction $(1 - 1/e) \approx 63.2\%$ of the unique examples of D, the rest being duplicates.

1.2 Random Forests

Random Forests is a substantial modification of Bagging that builds a large collection of de-correlated trees, and then averages them. When building these decision trees, each time a split in a tree is considered, a random sample of k features is chosen as split candidates from the full set of features. The split is allowed to use only one of those k features. A fresh sample of k features is taken at each node, and the best feature/split-point among the k is chosen.

For classification problems, $k = \sqrt{p}$ is typically taken, where p is the number of features in the model. For regression problems, the inventors recommend $k = p/3$, with a minimum node size of 5 as the default.

In Random Forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

1. Each tree is constructed using a different bootstrap sample from the original data.
2. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k-th tree.

3. Let $y_{\text{tree k}}(x)$ be the class prediction of the k-th Random Forest tree for x. Then the Random Forest classifier does a majority vote over all trees:

$$y_{\text{RF}}(x) = \text{majority}\{y_{\text{tree k}}(x), k = 1, \ldots\}$$

This is a draft chapter from the Kontinua Project. Please see our website (`https://kontinua.org/`) for more details.

Appendix A

Answers to Exercises

Index