Bagging and Random Forests

DS 301

Iowa State University

Ensemble method

- An ensemble method is an approach to combines many simple models in order to obtain a single and potentially very powerful model.
- These simple models are sometimes known as *weak learners*, since they may lead to mediocre results on their own.
- Ensemble methods for trees: bagging, random forests, boosting.

Bagging

- The decision trees we've covered so far suffer from *high variance*.
    - Conceptually this means, small changes in the training set can lead to large changes in the train.
- So how can we reduce the variance of the tree?

Stats Review

Suppose we are given a set of n independent observations $Z_1, \ldots, Z_n$. Each has their own variance σ^2.

- Variance of Z_i?
- Variance of $\bar{Z}$?

Averaging a set of observations reduces variance!

Apply this logic to trees

- How to reduce variance in trees?
- Take the average of them!
- Idea: take many training sets from the population, construct a tree using each training set, and average the resulting predictions:

$$\hat{f}_{\text{avg}(x)} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x).$$

Obvious problem with this approach?

Bootstrap

- Generate B different bootstrapped training sets.
- For the bth bootstrapped training set, we get a training and its predictions $\hat{f}^{(b)}(x)$.
- After we repeat this for all B of our bootstrapped training sets, we average all the predictions to obtain:

$$\hat{f}_{\text{bag}(x)} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{b}(x).$$

- This is called bagging.

Bagging

To apply bagging to regression trees:

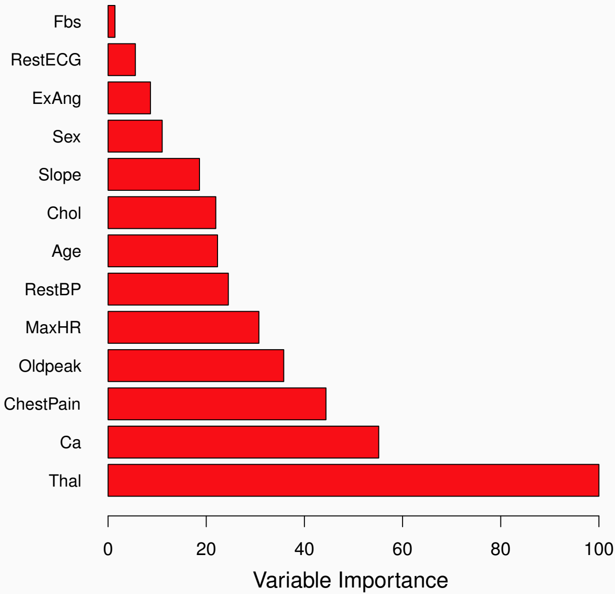
- Construct B regression trees using B bootstrapped training sets, and average the resulting predictions.
- Each individual tree grown deep and not pruned.
- Each tree has high variance, but low bias.
- Averaging these B trees reduces the variance.
- Improvements in accuracy by combining together hundreds or even thousands of trees into a single procedure.
- Number of B is not critical with bagging. If B is very large, it will not lead to overfitting.

Variable Important Measures

Bagging improves prediction accuracy at the expense of interpretability. How do we interpret the resulting tree?

- Overall summary of the importance of each prediction using the RSS (regression trees) or Gini index (classification trees).
- Regression trees: record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all B trees. A large value indicates an important predictor.
- Classification: record the total amount the Gini index is decreased by splits over a given predictor, averaged over all B trees.

Example - heart dataset



Problem with bagging

Random Forests

Random forests provide an improvement over bagged trees by *decorrelating* trees.

- Each time a split in a tree is considered, a *random sample of m predictors* is chosen as candidates from the full set of p predictors.
- Split is only allowed to use those m predictors.
- A new sample of m predictors is taken at each split.
- Typically we choose $m \approx \sqrt{p}$.
- This process of decorrelating the trees has the effect of reducing the variance.