Bagging (Bootstrap Aggregation)

- Training data: $\mathbf{Z} = \{(x_i, y_i)_{i=1}^n\}$
- Bootstrap samples[a]: $\mathbf{Z}^{*b} = \{(x_i^{*b}, y_i^{*b})_{i=1}^n\}$, where $b = 1 : B$
- $\hat{f}^{*b}$: classification/regression function trained by $\mathbf{Z}^{*b}$
- The bagging estimate is defined to be

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

- Advantage: reduce variance. So works well for high-variance, low-bias procedures, such as trees.

[a] sample with replacement from $\mathbf{Z}$.

Random Forest

1. For $b = 1 : B$:
   (a) Draw a bs sample $\mathbf{Z}^{*b}$ from the training data.
   (b) Grow a BIG tree T_b (with some restriction).
2. Output the forest $\{T_b\}_{b=1}^{B}$.

To make a prediction at a new point x

Regression: $\frac{1}{B}\sum T_b(x)$.

Classification: majority voting among $T_1(x), \ldots, T_B(x)$.

Restriction when growing a tree in the forest:

- At each split, randomly select m variables from the p variables, and then pick the best split among them.
- The recommended value for m is $\sqrt{p}$ for classification and $p/3$ for regression.
- Purpose: reduce the correlation between trees in the forest.

Out-of-Bag (OOB) Samples

- OOB samples: sample points which are not included in $\mathbf{Z}^{*b}$, i.e., they are not used in building the tree T_b

- The OOB samples can be used to get a test error for T_b.

- The prediction and error rate returned by randomForest are calculated based on OOB. The error is usually close to a CV error.

Variable Importance

- Calculation of the m-th variable's importance based on Gini Index: the improvement in the split-criterion (Gini index) is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable.

- The calculation can be easily extended to regression trees based on MSE.

- Another measure is computed from permuting OOB samples: For each tree T_b in the forest, calculate the prediction error (error rate for classification, MSE for regression). Then the same is done after permuting the jth predictor for the OOB samples. The difference between the two (before and after permutation) is then averaged over all trees, and normalized by the corresponding standard deviation[a].
- R returns both scaled and unscaled variable importance.

[a] If the standard deviation of the differences is equal to 0 for a variable, then the division is not applied.