Lecture 16: Random Forests
STAT 452, Spring 2021

- Decision trees are simple and useful for interpretation.
- However they typically are not competitive with the best statistical learning methods in terms of prediction accuracy.
- Random forests is a popular method that involves fitting a large number of decision trees which are then combined to yield a single consensus prediction.
- The random forests approach to combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss in interpretation.

Bootstraping

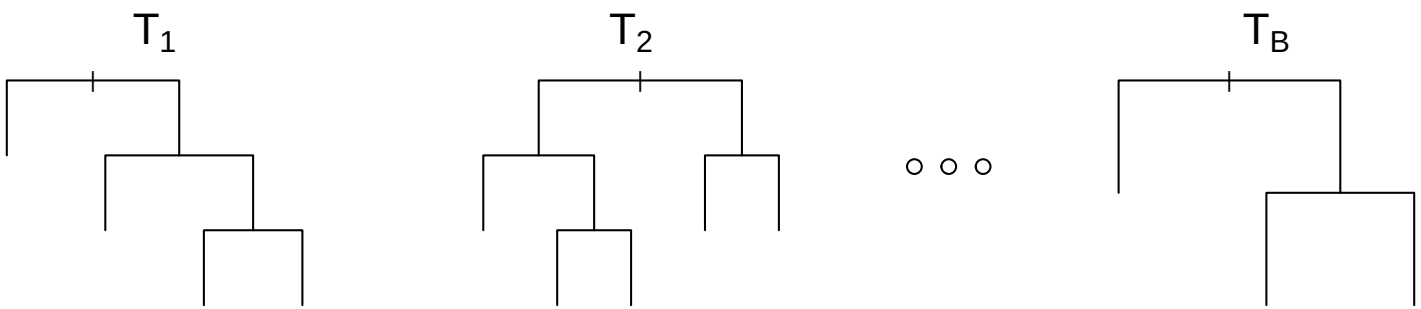
How can we estimate and then combine multiple tree models when we only have one training set?

- We can accomplish this by using a technique called *the bootstrap*.
- The bootstrap is a procedure that creates *replicates* of the original data set by randomly sampling the rows with replacement. So in a bootstrap dataset some rows will be repeated.
- For example suppose we have a data set with 10 rows numbered $1, 2, 3, ..., 10$. Then the following are bootstrap samples of those rows, which were generated using the R function `sample()`:

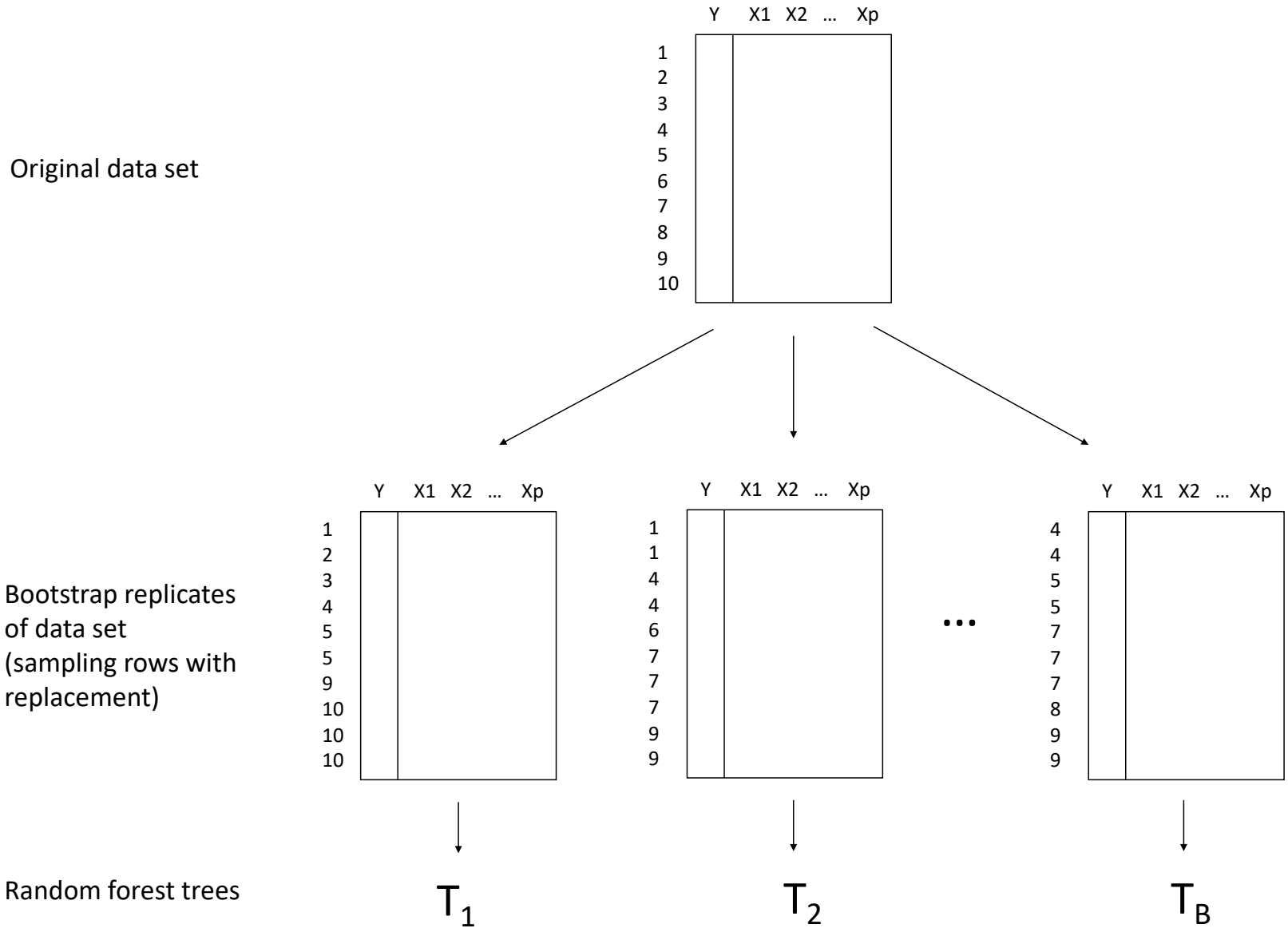
```
> sample(1:10, size = 10, replace = TRUE)
 [1] 10  7  1  2  3  1  9  5 10  5

> sample(1:10, size = 10, replace = TRUE)
 [1]  2  3  8  2  6  4 10  8  7  9
```

Random Forests

A random forest (RF) model is a collection of decision trees $\{T_b : b = 1, \cdots, B\}$ built from bootstrap samples of the data set.



Additionally, at each internal node of each tree a subset of $m \leq p$ predictors are randomly sampled as candidates for splitting.


Original data set
Y X1 X2 ... Xp
1
2
3
4
5
6
7
8
9
10
Bootstrap replicates of data set (sampling rows with replacement)
Y X1 X2 ... Xp
1
2
3
4
5
5
9
10
10
10
Y X1 X2 ... Xp
1
1
4
4
6
7
7
7
9
9
...
Y X1 X2 ... Xp
4
4
5
5
7
7
7
8
9
9
Random forest trees
T_1
T_2
T_B

Prediction

Let $\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 \cdots x_p \end{pmatrix}$ be a vector of new values for the predictors, and $T_b(\boldsymbol{x})$ the response prediction of the b^{th} tree.

The random forest prediction for a quantitative response is found by averaging over the predictions made by each tree in the ensemble:

$$\hat{f}_{RF}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\boldsymbol{x})$$

Prediction

How can random forests be used for classification?

For classification, $T_b(\boldsymbol{x})$ would give the class prediction for the b^{th} tree.

Then the random forest prediction is given by the *majority vote* of $\{T_b(\boldsymbol{x}) : b = 1, \cdots, B\}$. That is, the most commonly occurring class among the B predictions.

Tuning

RF has two main tuning parameters:

`mtry`: Number of predictors randomly sampled as candidates at each split.

- Has the effect of decorrelating, or diverisfying, the trees in the ensemble.
- Some special cases:
  - `mtry` $= p$ is called *bagging* (all the predictors are considered at each split)
  - `mtry` $= p/3$ is the default for regression
  - `mtry` $= \sqrt{p}$ is the default for classification

`ntree`: Number of trees; the default is 500.

RF is generally insensitive to the choice of these tuning parameters. The defaults work adequately well for most data sets.

Out-of-Bag Error

- The trees in a random forest model are fit to bootstrap samples of the data set.
- One can show that on average, each tree in a random forest model makes use of about two-thirds of the observations.
- The remaining one-third of the observations not used to fit a random forest tree are referred to as the *out-of-bag* (OOB) data.
- We can predict the response for the i^{th} observation using each of the trees in which that observation was OOB. This will yield around $B/3$ predictions for the i^{th} observation, which are then averaged.
- Thus we can get predictions for all n observations using the OOB data. The predictions can be used to calculate an essentially cross-validated MSE or classification accuracy.

Variable Importance

- Although the collection of random forest trees is much more difficult to interpret than a single tree, one can use the OOB data to obtain an overall summary of the importance of each predictor.
- One variable importance (VI) measure is computed by permuting the OOB data. This VI measure for predictor $j \in 1, \cdots, p$ is computed in the following way:
  - For each tree, the prediction error on the OOB data is recorded (error rate for classification, MSE for regression).
  - Then the prediction error for each tree is recorded again after permuting the values for the j^{th} predictor variable in the OOB data.
  - The difference between the two errors are then averaged over all the trees.

Concluding Remarks

- RF often performs well since it can model nonlinear relationships and high-order interactions between the predictors and response variable.
- RF is also robust to the inclusion of many predictors (i.e., large p). RF can be used when there are hundreds, or even thousands, of features without overfitting the data.
- One shortcoming of random forests is that it is a so-called "black-box" method. In contrast to linear regression, or individual decision trees, we do not know what the relationships are between the response and predictor variables. However, variable importance plots allow for some interpretation with random forests.