Tree-based regression methods

- Tree-based methods can be applied for continuous (Regression Trees) or categorical (Classification Trees) Variables
- The method consists in stratifying or dividing the predictors space in a set of simple segments
- In order to make predictions for a given observation we normally take the mean or the mode of the training observations located in the region for which that observation belongs to.

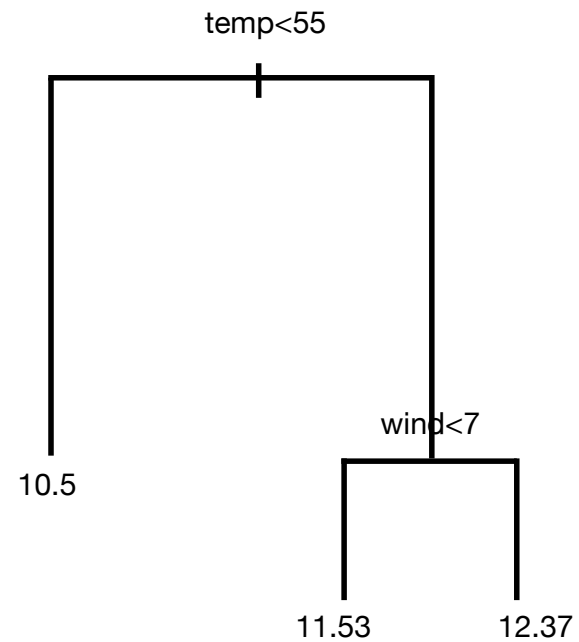
- The set of rules used to segment the predictor space can be represented in a tree. They are normally called decision trees.
- The tree stratifies or segments the observations in regions involving the predictors.
- The regions $R_1, R_2, \ldots, R_J$ are known as the *leaves or terminal nodes* of a tree.
- The decision trees are normally drawn upside down, so the leaves are at the bottom of the tree.

Ozone data set

- The tree *has internal nodes*. These are the points where the predictor space is split. For example, $temp < 55$ splits the predictor space in two regions according to variable $temp$. To the right you will have observations corresponding to the values $temp \geq 55$, and to the left of the tree, observations corresponding to the values $temp < 55$
- The right branch is split in two branches according to the value of wind.
- The three values in the final leaves or terminal nodes are the mean values of ozone in the different regions.

Ozone Data set: Decision tree example

- Two internal nodes
- Three leaves (terminal nodes)



Process of building a Regression tree

- Divide the predictor space into J non-overlapping regions (rectangles): $R_1, R_2, \ldots, R_J$, in order to minimize the Residual Sum of Squares (RSS):

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

  where $\hat{y}_{R_j}$ is the mean response of the training observation in box j.
- Make the same prediction for all observations that fall in region R_j. This is the mean response value of the training observations.
- Use recursive binary splitting (top-down approach). Each split has two branches. Process continues until a stopping criteria is reached. (For example: each terminal node or leaf has lower than a required minimum number of observations).

Tree pruning

- The tree building process might overfit the data and the tree might be too complex.
- Best strategy: Grow a large tree T_0 and prune it back to obtain a sub-tree.
- Cost complexity pruning (CC): Consider a sequence of trees indexed by a tuning parameter α.
  - For each value of α there is a sub-tree $T \subset T_0$ that minimizes:

$$CC(T) = RSS(T) + \alpha|T|$$

    where $|T|$ is the size of the tree (number of final leaves or terminal nodes).

- α controls tree complexity versus the fit to the training data.
- If $\alpha = 0$, $T = T_0$ which is the original tree.
- When α increases from zero, branches of the tree are pruned in a nested way. The larger the value of α the larger the cost to have a more complex tree.
- Regression trees are fitted using *rpart* library. It assumes a default value for α (the complexity parameter $cp = 0.01$). Many of the tree building parameters are controlled by the function *rpart.control*, including the stopping rules for tree building:

```
rpart.control(minsplit = 20, minbucket = round(minsplit/3),
cp = 0.01, maxcompete = 4, maxsurrogate = 5,
usesurrogate = 2, xval = 10, surrogatestyle = 0,
maxdepth = 30, ...)
```

Classification Trees

- Classification trees are similar to regression trees but are used to predict qualitative responses.
- Predictions for observations in a region are the most commonly occurring class for the training observations in the region.
- The RSS is not suitable for the recursive binary splitting. Different measures of the classification error rate are used instead.

Splitting measures

If you have K classes, $\hat{p}_{jk} = \frac{n_{jk}}{n_j}$ represents the proportion of observations in class k at node j, where $k = 1, \ldots, K$

Gini Index

It is a measure of the total variance among the K classes at node j:

$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$$

It is also a measure of node purity.

Classification Error

$$E = 1 - \max_k \hat{p}_{jk}$$

Entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{jk} \log(\hat{p}_{jk})$$

The Deviance can also be used as a splitting measure. The entropy and Gini index are similar numerically.

Random Forests

- Regression or Classification trees may not have the same level of accuracy for prediction as the standard regression methods.
- They might be very sensitive to a small data change.
- Their predictive performance can be improved by aggregating many trees using different methods. One possible method is to use *Random Forests.*

- Take many samples (*bootstrapped samples*) from the training data set to build decision trees for each sample.
- For each possible node split, consider a size m random sample of the original p predictors as possible split candidates.
- The split uses only one of the m selected predictors.
- Each split uses a fresh sample of m candidate predictors with $m \approx \sqrt{p}$.
- This process *decorrelates* the successive trees.
- If a random forest is built by choosing $m = p$ predictors as candidates at each split, the process is called *bagging*.
- The final prediction results are obtained from the average of all trees.

- Observations not used to fit a tree are call out-of-bag (OOB) observations.
- We can predict the response of a given observation, using the trees in which that observation was an OOB observation.
- We can take the average of the predictions (in regression trees) or the most common class (in classification trees) to get a single prediction for each observation.
- The resulting OBB MSE error (regression) or classification error is an estimate of the *test error*. This is more convenient than doing cross-validation, specially for large data sets.