Random Forests

Random forests

- Random forests is a specific technique that applies bootstrap-and-aggregate or *bagging* using *CART* as base classifiers.
- Random forests builds a large collection of de-correlated trees and then aggregates them.
- Their performance is on par with boosting and they are very easy to tune.
- As a consequence, they are very popular and implemented in many packages.

Random forests

1. For m = 1 to M (number of trees in the forests) :
   a. Draw a bootstrap sample of size N_m from the training data $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$
   b. Grow a random-forest tree T_m to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until a minimum node size n_{min} is reached:
      i. Select d variables from the D variables
      ii. Pick the best variable/split-point among the d
      iii. Split the node into two child nodes
2. Return the ensemble of trees $\{T_m\}_1^B$

Random forests

To make a prediction at a new point **x** :

- Regression:

$$y_B(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} T_m(\mathbf{x})$$

- Classification: Use majority vote.

Random forests

Interesting ‘extra’ features that random forests give

- Out of bag samples : cross validation comes for free.
- Variable importance : It is easy to calculate how important each variable in your data is.
- Proximity plots : Visualization of how close points are to one another in the training set.
- Robust to over-fitting.