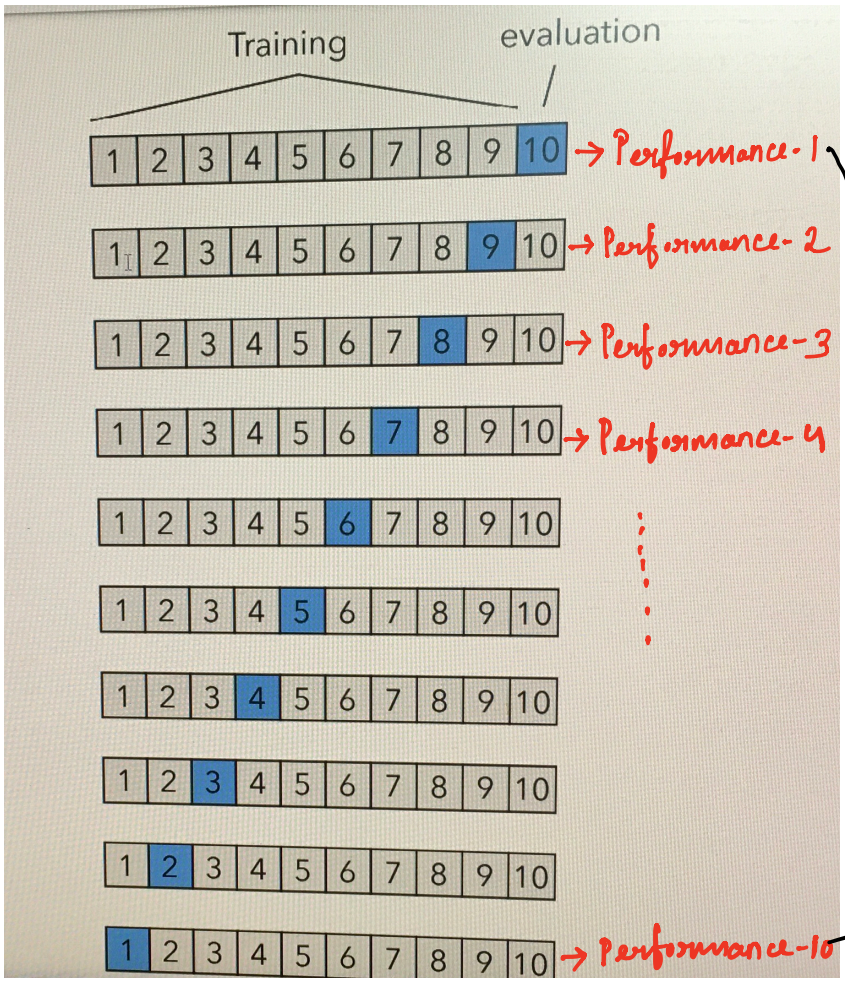
Cross Validation



Bootstrap Sampling

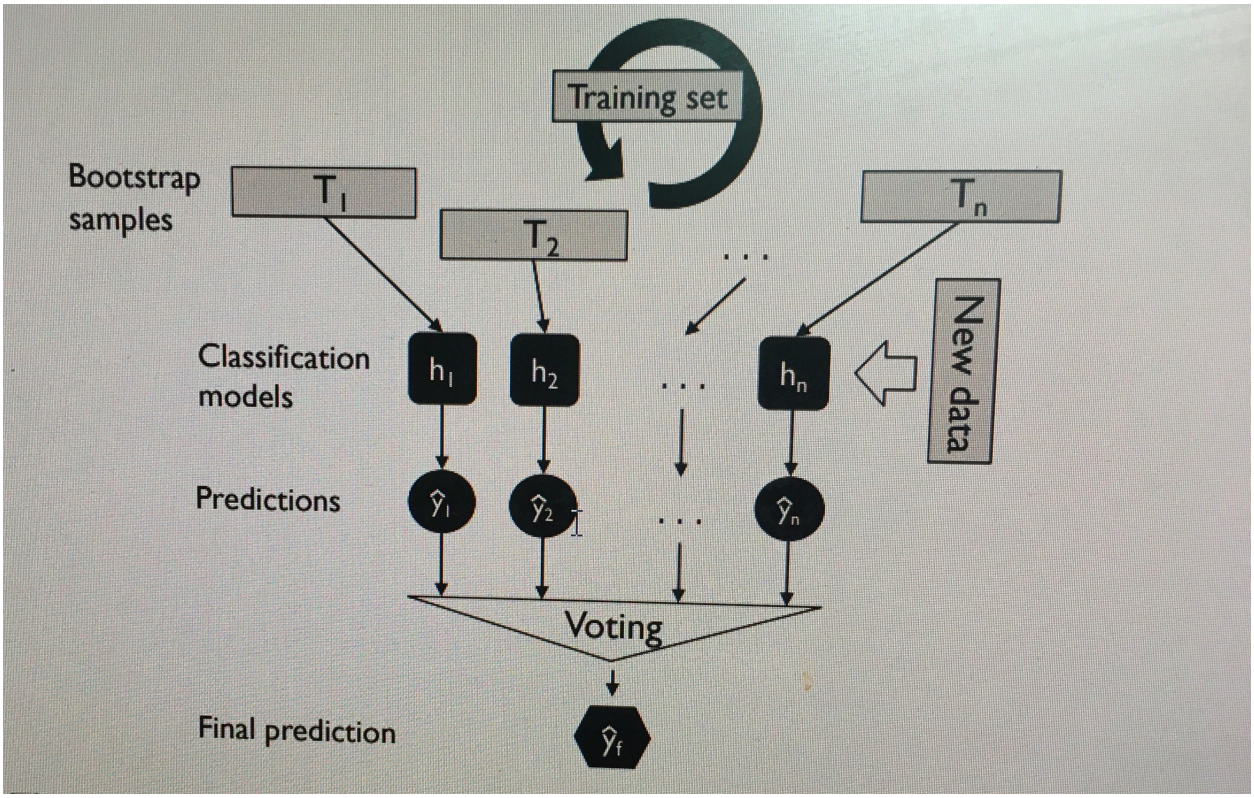
- Original data (random sample):
  - {1,2,3,1,2,1,1,1} (n = 8)
- Bootstrap random sample data:
  - {1,2,1,2,1,1,3} (n = 8); mean = 1.375
  - {1,1,2,2,3,1,1} (n = 8); mean = 1.375
  - {1,2,1,2,1,1,2} (n = 8); mean = 1.25

Estimated Mean ↓

(1.375 + 1.375 + 1.25) / 3

≃ 1.333

Bagging


Training set
Bootstrap samples
T_1
T_2
...
T_n
Classification models
h_1
h_2
...
h_n
New data
Predictions
$\hat{y}_1$
$\hat{y}_2$
...
$\hat{y}_n$
Voting
Final prediction
$\hat{y}_f$

Random Forest

↓

Combination of uncorrelated DTs

↓

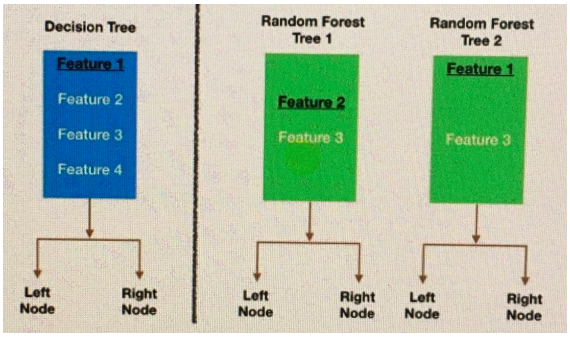
How does Random forest ensures that each individual trees is not too correlated?

↙ ↘

① Bagging (Bootstrap Aggegation)

② Feature Randomness

Allow each individual tree to randomly sample from the dataset with replacement, resulting in different trees.



↓

Randomly select subset of features

Parameters

n_estimators, default=100

The number of trees in the forest.

max_depth, default=None

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_split or float, default=2

The minimum number of samples required to split an internal node:

min_samples_leafint or float, default=1

The minimum number of samples required to be at a leaf node.

max_features{"auto", "sqrt", "log2"}, int or float, default="auto"

The number of features to consider when looking for the best split:

If int, then consider max_features features at each split.

If float, then max_features is a fraction and int(max_features * n_features) features are considered at each split.
If "auto", then max_features=sqrt(n_features).
If "sqrt", then max_features=sqrt(n_features) (same as "auto").
If "log2", then max_features=log2(n_features).
If None, then max_features=n_features.
Note: the search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.

max_leaf_nodes, default=None

Grow trees with max_leaf_nodes in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes.