TUTORIAL 4

ELENA MARÍA RUIZ IZQUIERDO
STEFANOS MANDALAS

Ensemble Learning

DEFINITION

- Ensembles are a divide-and-conquer approach used to improve performance. The main principle behind ensemble methods is that a group of weak learners can come together to form a strong learner.
  - Combines the results from different models.
  - Models can be a similar type or different.
  - The result from an ensemble model is usually better than the result from one of the individual models.

- Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one.

BAGGING METHODS

- Estimators of same type
- Built independently
- Ensemble = average of predictions
- Variance is reduced

Random Subspaces

Negative Correlation

RandomForest

BOOSTING METHODS

- Estimators of same type
- Built sequentially
- Bias is reduced in each step

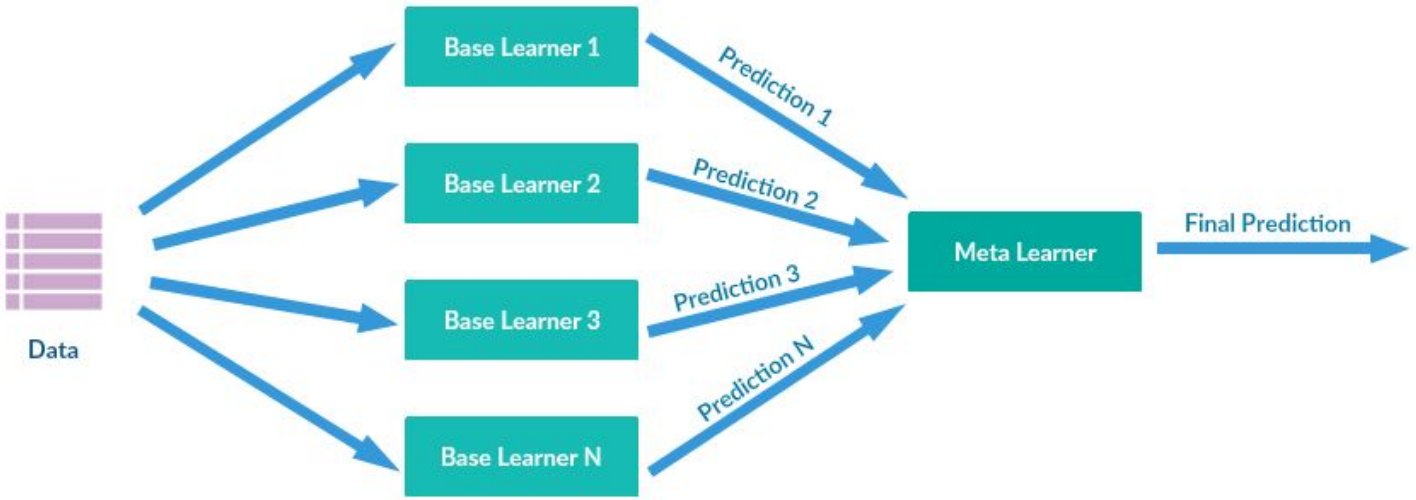
AdaBoost

Gradient Tree

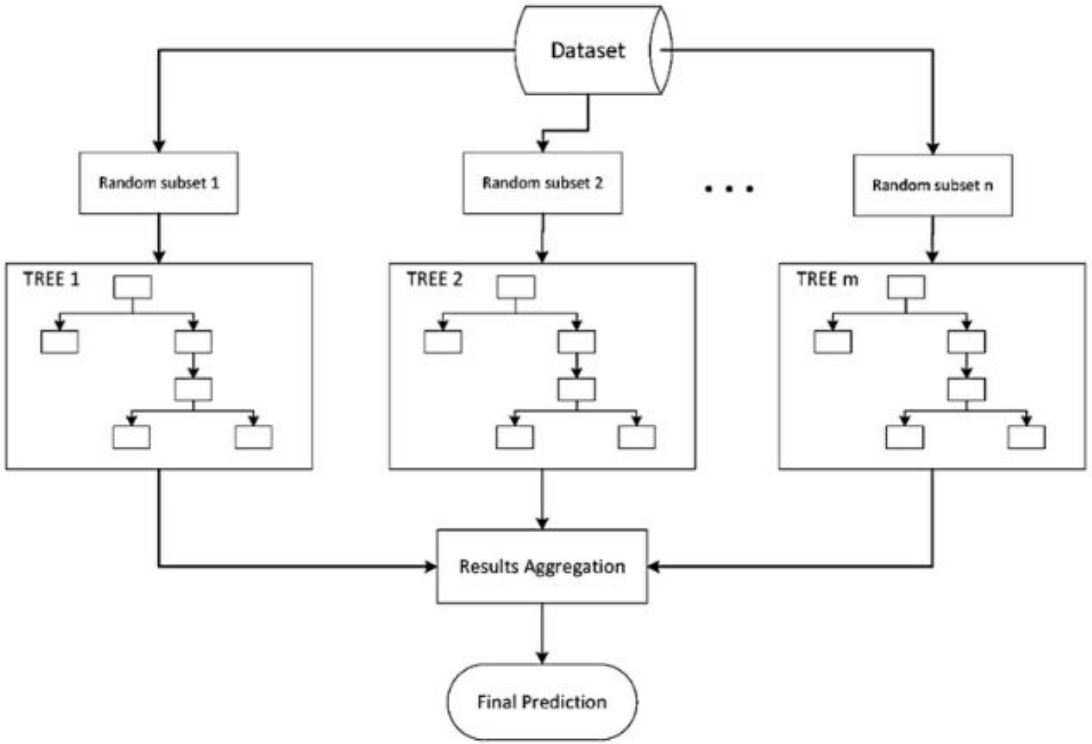
Histogram-based

STACKING METHODS

- Estimators of different type
- Built in parallel
- Predictions are used to train a “meta model”
- Bias is reduced

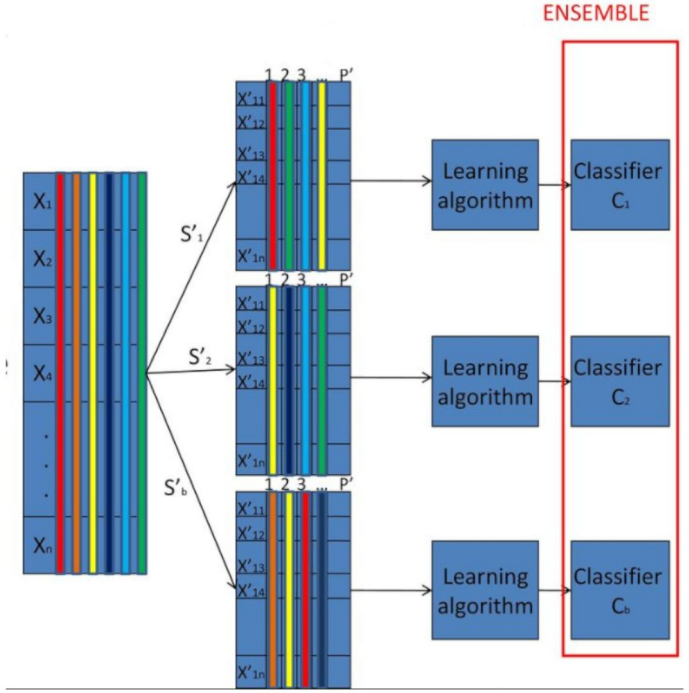


BAGGING



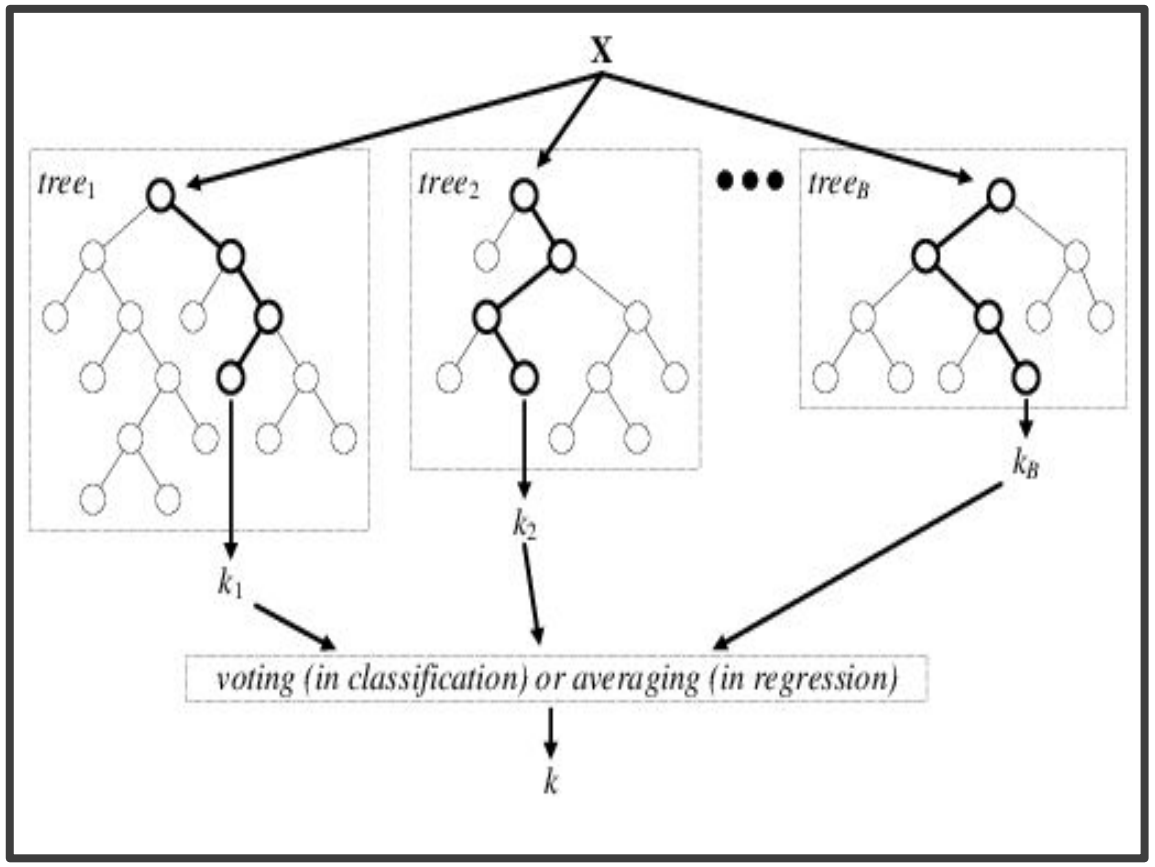
- Bootstrap Aggregating
- Given a sample of data, multiple bootstrapped subsamples are pulled. A Decision Tree is formed on each of the bootstrapped subsamples. After each subsample Decision Tree has been formed, an algorithm is used to aggregate over the Decision Trees to form the most efficient predictor.

RANDOM SUBSPACE



- Is an ensemble learning method that attempts to reduce the correlation between estimators in an ensemble by training them on random samples of features instead of the entire feature set.
- Random subspaces are an attractive choice for problems where the number of features is much larger than the number of training points

RANDOM FOREST



- An ensemble classifier using many decision tree models.
- Can be used for classification or regression.
- Trees are weak learners and the random forest is a strong learner.
- How random forest work:
  - A different subset of the training data are selected (2/3), with replacement, to train each tree.
  - Class assignment is made by the number of votes from all the trees and for regression the average of the results is used.
- Use a subset of variables:
  - A randomly selected subset of variables is used to split each node.

ADABOOST

- Adaptive Boosting
- Extends boosting to multi-class and regression problems.
- It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.
- Adaboost helps you combine multiple “weak classifiers” into a single “strong classifier”.