Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Random Forest & Extremely Randomized Trees (Extra Trees)

Alessia Bisio
Iratxe Campo
Ander Carrera
Asier De La Natividad
Jon Ander Sukia

Index

Mondragon Unibertsitatea
Faculty of Engineering

1. Description

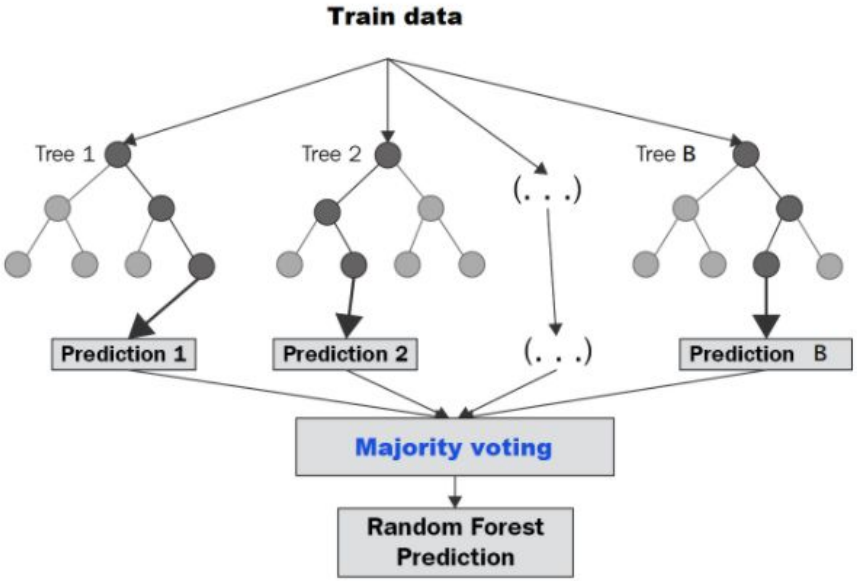
Description: Random Forest

Mondragon Unibertsitatea
Faculty of Engineering

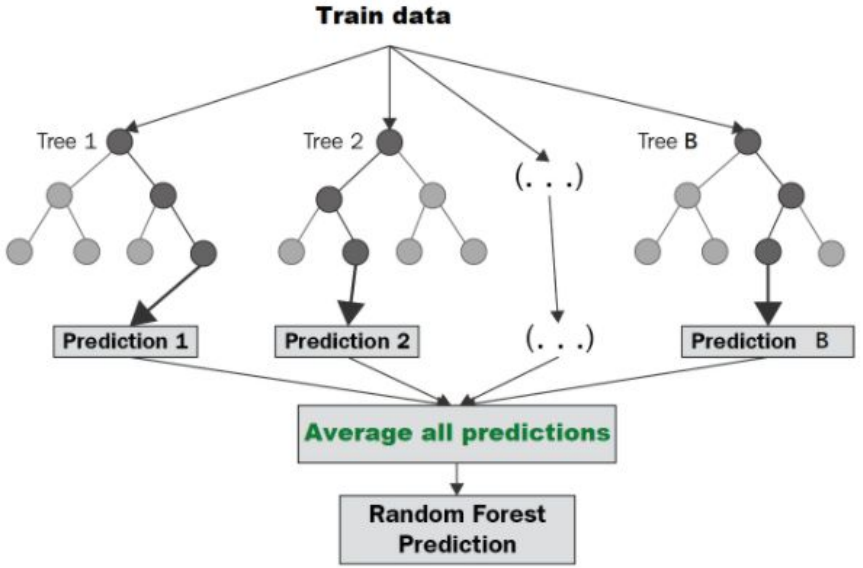
- One of the most used machine learning algorithms.
- Useful for classification and regression problems.
- It is a supervised learning algorithm that is composed by decision trees (forest).
- Usually trained with bagging method.
- Adds additional randomness to the model. Searches for the best feature among a random subset of features.

Description: Random Forest

Mondragon Unibertsitatea
Faculty of Engineering



For classification



For regression

Description: Extremely Randomized Trees (Extra Trees)

Mondragon Unibertsitatea
Faculty of Engineering

- Algorithm related to other ensembles of decision trees algorithms (Bagging and Random Forest)
- More randomness during the training phase → improved variance/bias trade-off
- Random split at each node → Faster (10 times) and less computationally expensive than Random Forest
- All the data available in the training set are used to build each stump
- Predictions made in the **same way of Random Forest for both Classification and Regression problems**.
- It can be used with Regression as well

Description: Extremely Randomized Trees (Extra Trees)

Mondragon Unibertsitatea
Faculty of Engineering

	Decision Tree	Random Forest	Extra Trees
Number of trees	1	Many	Many
No of features considered for split at each decision node	All Features	Random subset of Features	Random subset of Features
Boostrapping(Drawing Sampling without replacement)	Not applied	Yes	No
How split is made	Best Split	Best Split	Random Split

	Extra Trees		Random Forest	
	F-Score	Time(s)	F-Score	Time(s)
Dataset 1	0.773 ± 0.009	0.213	0.778 ± 0.029	0.345
Dataset 2	0.775 ± 0.027	0.206	0.784 ± 0.030	0.349
Dataset 3	0.807 ± 0.015	0.206	0.811 ± 0.019	0.291
Dataset 4	0.755 ± 0.016	0.238	0.788 ± 0.023	0.283
Dataset 5	0.758 ± 0.017	0.182	0.782 ± 0.022	0.357

Example: comparison between Random Forest and Extra Trees, results obtained using 5 different datasets.

2. Scikit-Learn function

Scikit-Learn function Random Forest

Mondragon Unibertsitatea
Faculty of Engineering

sklearn.ensemble.RandomForestClassifier

- Fits N decision tree classifiers on various sub-samples of the dataset. It uses averaging to improve the predictive accuracy and try to avoid overfitting.

Scikit-Learn function Random Forest

Mondragon Unibertsitatea
Faculty of Engineering

Parameters:

- `n_estimators` : Number of trees in the forest (Default = 100)
- `max_depth` : Maximum depth of the tree. If None → nodes are expanded until all leaves are pure.
- `min_samples_splitint`: The minimum number of samples required to split an internal node.

Scikit-Learn function Random Forest

Mondragon Unibertsitatea
Faculty of Engineering

- max_features:
  - auto → max_features=sqrt(n_features)(same as sqrt option)
  - log2 → max_features=log2(n_features)
  - None → max_features=n_features

- class_weight: {default: none} Weights associated with classes in the form {class_label: weight}. All classes are supposed to have weight 1 by default.

- bootstrap: (T/F), default=True
  - Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.

Scikit-Learn function Extra Trees

Mondragon Unibertsitatea
Faculty of Engineering

sklearn.ensemble.ExtraTreesClassifier

- This class creates an estimator that fits a number of randomized decision trees (extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Scikit-Learn function Extra Trees

Mondragon Unibertsitatea
Faculty of Engineering

Parameters:

- `bootstrap (True/False)`: Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.

- `max_samples (int/float/ default = None)`: If bootstrap is True, the number of samples to draw from X to train each base estimator.
  - `None : X.shape[0]`
  - `int: draw max_samples samples.`
  - `float: draw max_samples * X.shape[0] samples.`

- Extra trees also uses the same parameters mentioned before.

Other functions

Mondragon Unibertsitatea
Faculty of Engineering

- BaggingClassifier → In order to use bagging
- load_iris → In order to load an example dataset
- metrics → Extract confusion matrix and metrics to evaluate the model
- train_test_split → Split the dataset into train and test data

3. Demonstration

Demonstration

Mondragon Unibertsitatea
Faculty of Engineering

https://github.com/andercarrera/Extremely-Randomized-Trees-and-Random-Forest

4. Bibliography

Bibliography

- https://github.com/andercarrera/Extremely-Randomized-Trees-and-Random-Forest
- https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b
- https://towardsdatascience.com/tagged/extremely-randomized-tree?p=289ef991e723
- https://link.springer.com/content/pdf/10.1007/s10994-006-6226-1.pdf
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
- https://builtin.com/data-science/random-forest-algorithm
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/