

Random Forest

Pritam Prakash Shete

Computer Division, BARC

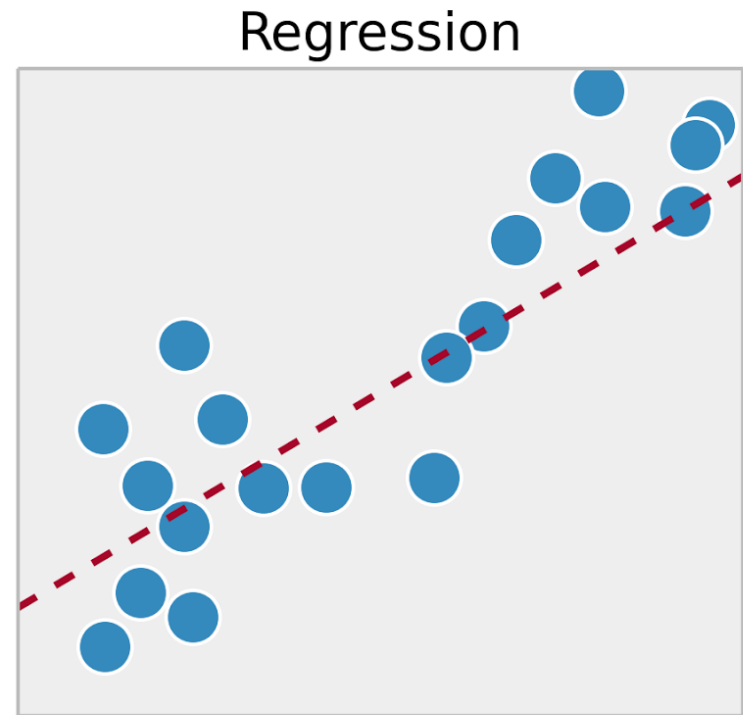
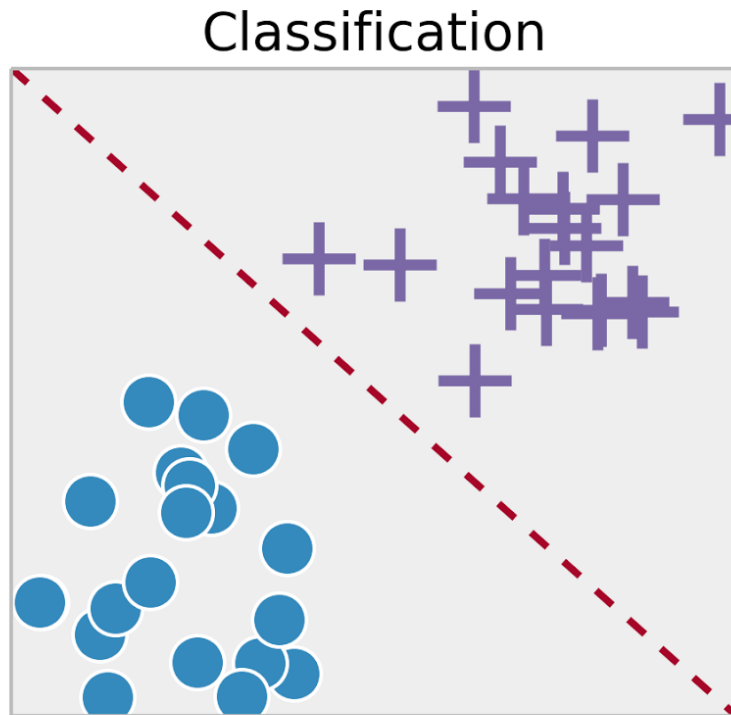
Centre for Excellence in Basic Sciences

Topics

- Supervised learning
- Decision tree
- Random forest
- Advantages
- Disadvantages

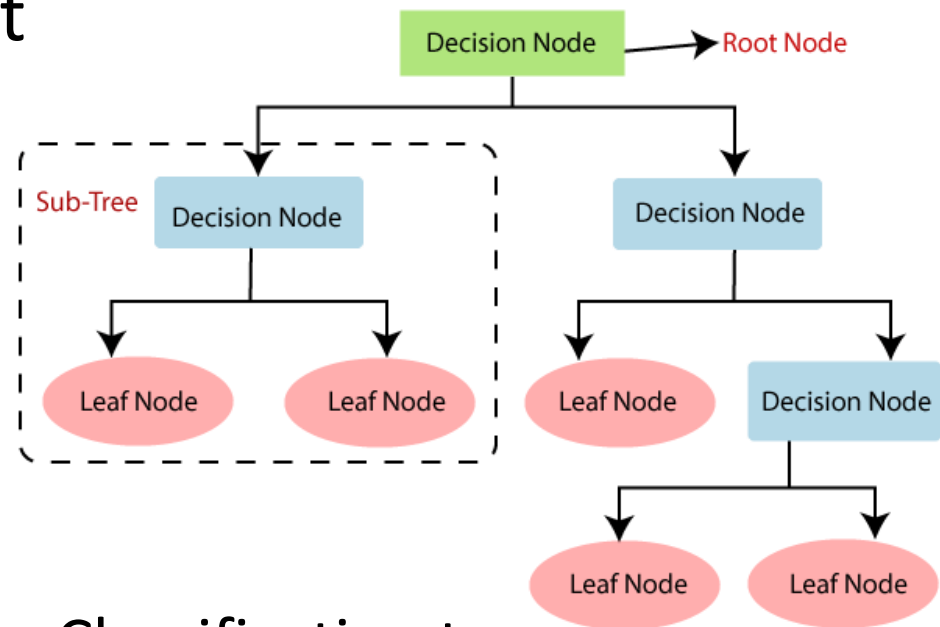
Supervised Learning

- Training set – $\{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)}) \}$
- Labeled dataset



Decision Tree

- Binary tree
- Recursively split dataset
- Root node
- Decision nodes
 - Split dataset
 - Select feature
 - Select split condition
- Pure leaf nodes
 - Class of dataset sample – Classification tree
 - Value of dataset sample – Regression tree



Decision Tree – Disadvantages

- Over-fitting
- Less accurate
- Solution – Use random forest algorithm

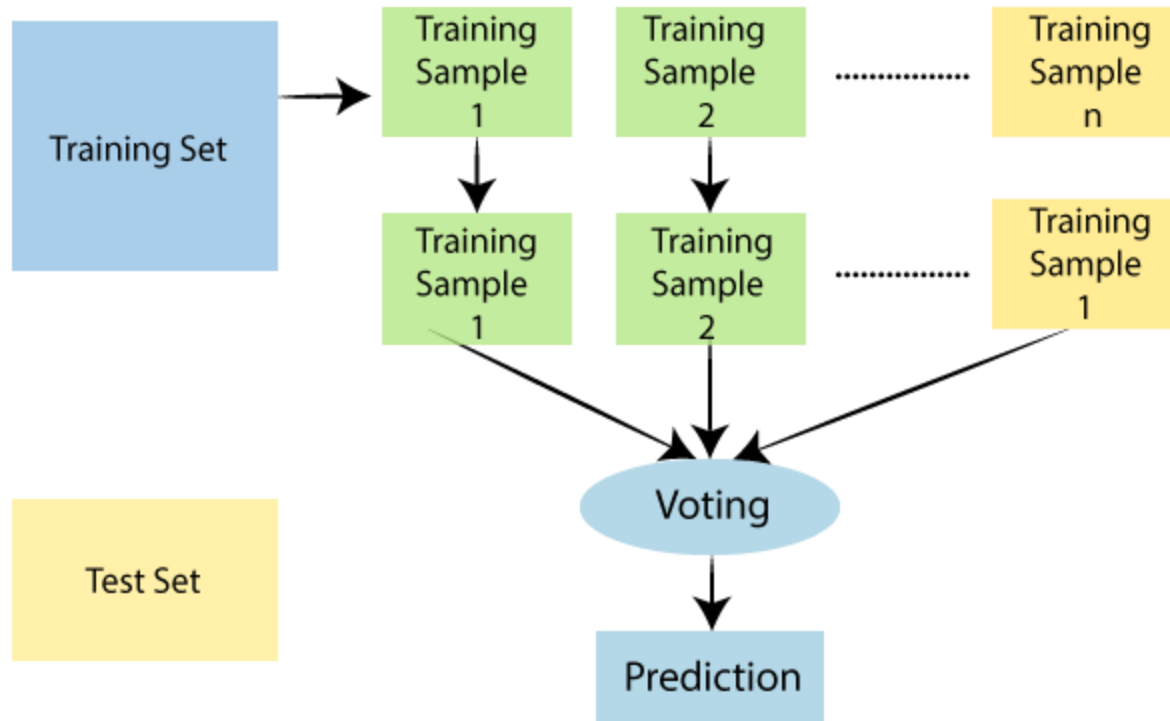
Random Forest

- Subset of dataset
 - Bootstrapped dataset
- Subset of features
 - Bagging approach
 - Ensemble learning
 - Improve performance
- Number of decision trees
 - More number
 - Less over-fitting
 - Higher accuracy

Random Forest Algorithm

1. Create bootstrapped dataset
2. Select subset of features
 - Bagging approach
3. Create decision tree using selected features
4. Repeat Steps 1 to 3
 - Number of decision trees

Random Forest



Advantages

- Less over-fitting by using multiple trees
- More accurate as compared to single tree

Disadvantages

- Computationally expensive
- Less interpretable than decision tree

Questions?

Thank you