Random Forest

Decision trees are easy to build & provide high interpretability

But They have high variance

They suffer from overfitting if allowed to grow without control.

They are not flexible enough for classifying new samples.

This is why a new model that has more flexibility is required

Random forest classifiers combine the simplicity of decision trees with flexibility

This results in vast improvement in accuracy

Random forest is a forest of decision trees.

"In diversity there is strength"

Random Forest

→ Random forest is extension of decision trees

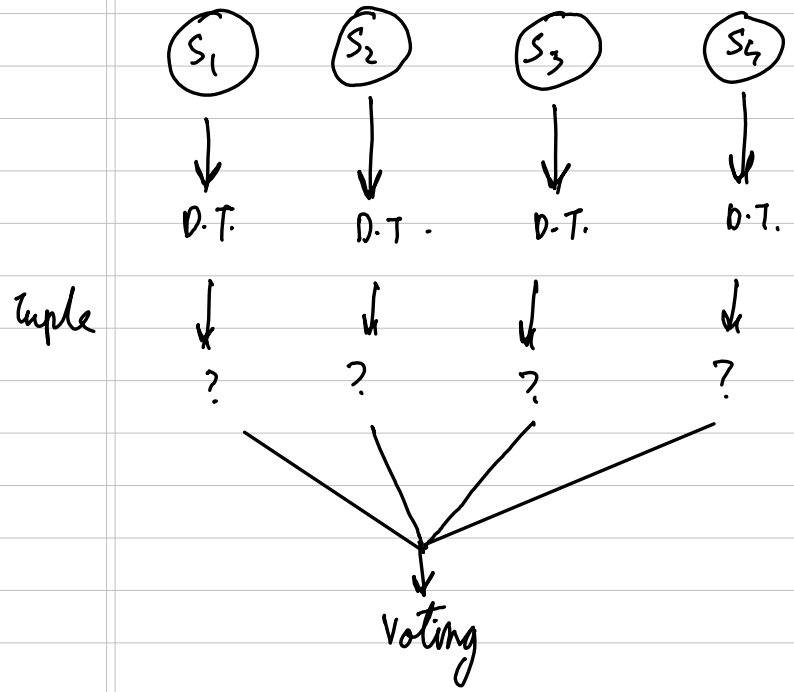
→ It selects random subsets of dataset

→ Then it applies decision tree on the subsets

→ Every tree predicts something

→ Then the majority votes are considered

→ Greater number of trees gives better trees



Training

step 1 : select random k data points

(This is known as bootstrapped dataset)

step 2 : Build decesion tree but with random feature subsets

step 3 : Repeat 2 & 3 N times

Output is a forest of trees

Testing

step 1 : Put tuple on every tree

step 2 : Take voting of all & choose maximum

"Majority Voting"

Bootstrapped dataset

→ Random rows (sample points) are selected

→ One sample can occur more than once

→ Columns (features) are preserved. All features are present for all points

---

→ After making bootstrap dataset, the trees can be made with this variation

→ Instead of considering all features at every step, select random selections of features at every step

→ This adds flexibility to the model.

→ Research tells that the ideal size of feature subset is

i) $\sqrt{N}$

ii) $\log(N)$

} N is total no of features

→ Random forest works on ensemble learning

→ Various models (here decision trees) are combined and made into one model

→ The technique of using bootstrapped dataset & majority voting is called "Bagging"

→ Two "random" processes are used

① Bootstrapping

② Random Feature selection

There is a seperate method in random forests to handel missing data

Advantages →

1. Can handel high dimentionality
2. Less overfitting
3. Large data handled (Large dimentions also)
4. Missing data handled ✓
5. Higher accuracy

Disadvantages

1. High time complexity
2. Less interpretable black boxes

Applications of random forest

1. Classification – spam detection, medical diagnosis
2. Regression – Predicting stock market, house price
3. Anomaly detection – fraud detection, outliers
4. Feature selection
5. Text classification – Sentiment analysis

Random forests are very versatile and useful algorithms. They are very Robust

Decision Tree	Random Forest
Suffer from overfitting	No overfitting
Faster	Slower
Makes rules based on data	Rules are made by underlying decision trees on random data samples
More Interpretable	less interpretable
Useful for small amount of features	Work even for large features
Error prone to noise	Robust from noise