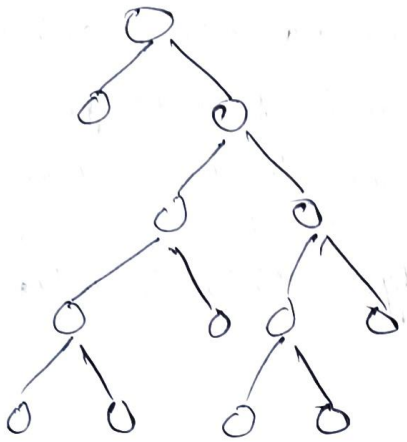
13-Nov-22

Decision Tree



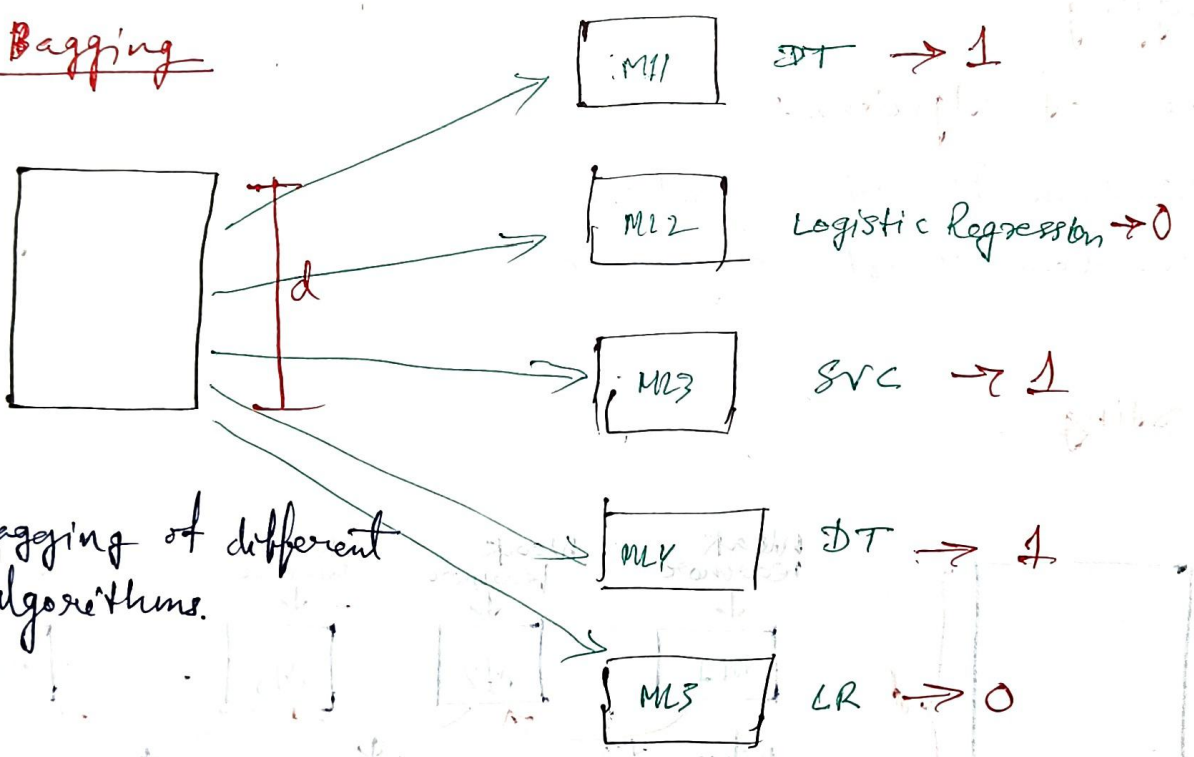
Overfitting issue

- Low Bias
- High Variance

} - Using Pre Pruning or Post Pruning we can avoid overfitting issue.
- Using bagging techniques also we can remove overfitting problem.

Bagging and Boosting

Bagging



Bagging of different algorithms.

> Final output will be based on majority voting classifier. Here it will be '1'.

> { Bootstrap → dividing into diff algorithms ; Aggregation → aggregate to get final result } - Ensemble Technique

> Each algorithms can learn something unique from the problem statement.

Example:

KBC competition,

Subhash → Data science domain knowledge.

X(UPSC) → different domain knowledge

So X(UPSC) candidate perform better than me at KBC competition. because he/she has different domain knowledge. This is called **Bagging.**

> Most of the time, Ensemble Techniques is a good choice.

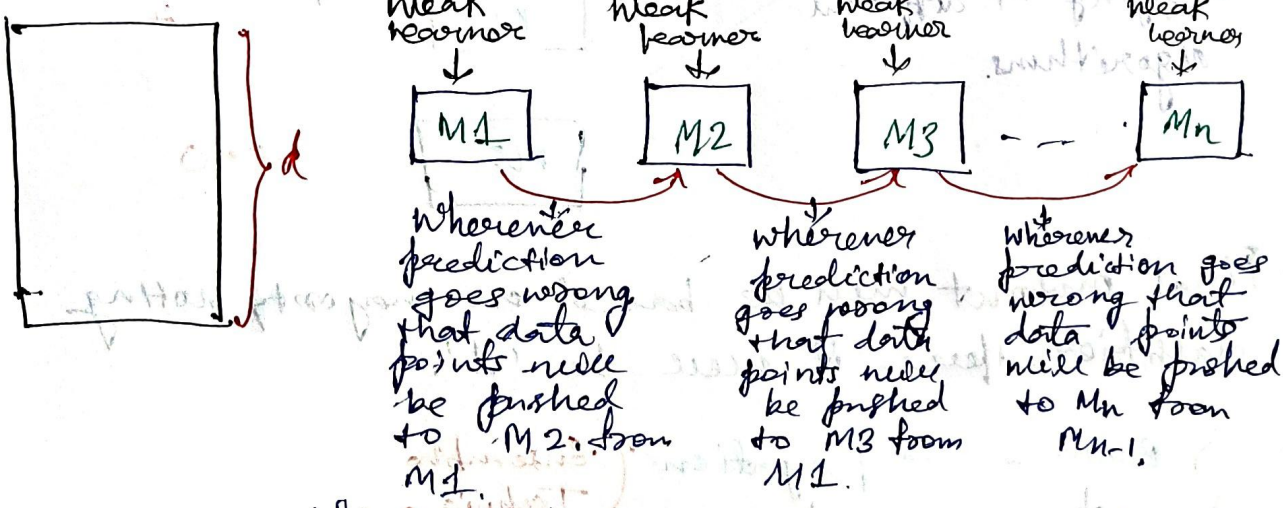
Bagging

Types of Algorithms

① Random Forest classifier

② Random Forest Regressor

Boosting



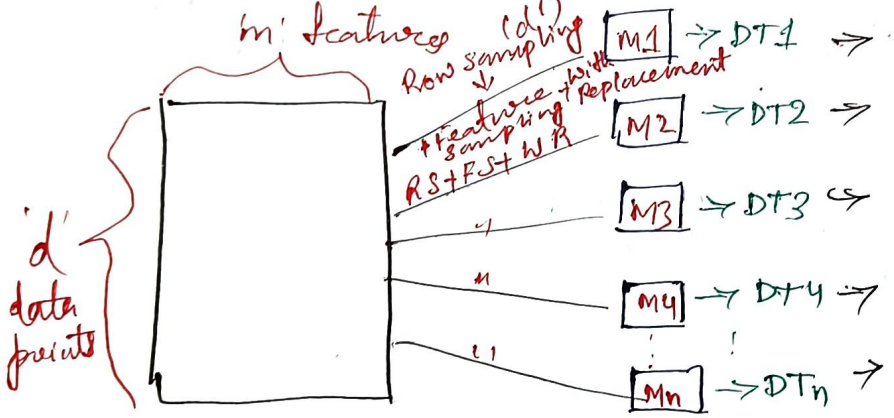
> After combining all the models, it will be strong learner.

Types of Algorithms

1. Adaboost Regressor and classifier
2. Gradient Boost Regressor and classifier
3. Xgboost Regressor and classifier
   ↓
   Extreme Gradient Boost

Random Forest classification and Regression

- It is an Ensemble Techniques.
  ↓
  combination of many models.



Classification
Majority output will be taken

Regression
Average will output will be taken

→ In sk-learn, we combine multiple decision trees

$d' < d$ $\quad$ $M1 < M$

For each model
{ Row sampling (d') + Feature Sampling (M1) + With Replacement }

Q Why are we using Random Forest?

Ans Pre-Pruning and Post-Pruning is a bigger task for large dataset. so, to avoid overfitting we use Random Forest. For huge dataset, we cannot see complete tree.

Overfitting

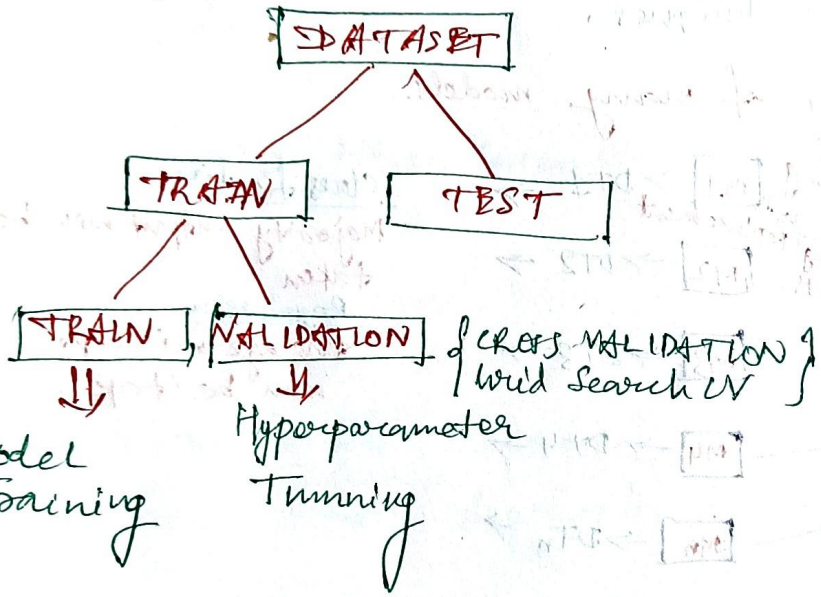
Low Bias
High Variance → connect using Random Forest → Low Variance

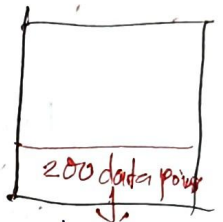
→ Using Random Forest, we can convert high variance to low variance.

Out of Bag Evaluation

⇒ while doing Row sampling, feature sampling and with replacement for each model, so there is chance that some set of values never get selected.

⇒ We can use unselected data points for testing and validation purpose.





Never get Selected because it never get chance

> If you set oob = true, then it will automatically handle missing data points for testing purpose.

7- ~~out of~~

Out of Bag Error (Validation Error)

$$\text{Out of Bag Error} = 1 - \text{Out of Bag score}$$

⇒ oob score is the accuracy score for the validation data.

WRT Regression

> R-2 score can be used.

> We decide algorithm based on the dataset.
> Random forest has more time and complexities.

> Bagging can be heterogeneous as well as homogeneous.

Hyperparameters

n_estimators = number of Decision Trees

b bootstrap = combination of models

oob_score = accuracy score, used left data for testing purpose.

n_jobs = How much CPU our model will take while running

~~verbose~~

verbose = can see messages about configuration while running

Booting classifier

- We can use different algorithms.