Random forest classifier

Data science offers a plethora of classification algo. such as logistic regression, SVM, naive Bayes and decision trees.

Individual decision trees can be combined to form a random forest.

core idea of decision trees: → what feature will allow me to split the observations at hand in a way that the resulting groups are as different from each other as possible (and the members of each resulting subgroup are as similar to each other).

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with most of the votes becomes our model's prediction.

wisdom of crowds:

A large no. of relatively uncorrelated models (trees) operating as a committe will outperform any of the individual constituent models.

The reason for this wonderful effect is that the trees protect each other from their individual errors.

Ensuring that the models diversify each other:

ie. the correlation between them should be the least.

① Bagging (Bootstrap Aggregation): Decision trees are very sensitive to the data they are trained on → small changes to the training set can result in significantly diff. tree structures.

Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement → this process is called bagging.

② Feature randomness:

In normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs right node.

Applications:

① Banking: RF is used to find loyal customers, those who can take out plenty of loans and pay interest to the bank properly.

② Medicine: RF can be used to identify both correct combination of components in medicines and to identify diseases by analyzing patient's medical records.

③ customer: predicting whether the customer will like the recommended product based on the experience of similar customers.

Advantages:

① For applications in classification problems, RF will avoid overfitting.

② same algo can be used for both regr / classif

③ RF can be used for finding the most import features → feature engineering.

Random forest is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Instead of relying on one decision tree, the RF algo. takes the prediction from each tree and based on the majority of votes it predicts the final output.

Greater no. of trees in a forest leads to higher accuracy and prevents the problem of overfitting.

Algo:

RF works in two-phases

① create the RF by combining N decision trees.

② make predictions for each tree created in phase1

① select random K data points from training set.

② Build the decision trees associated with the selected data points (subsets).

③ choose the no. of number (N) for decision trees that you want to build.

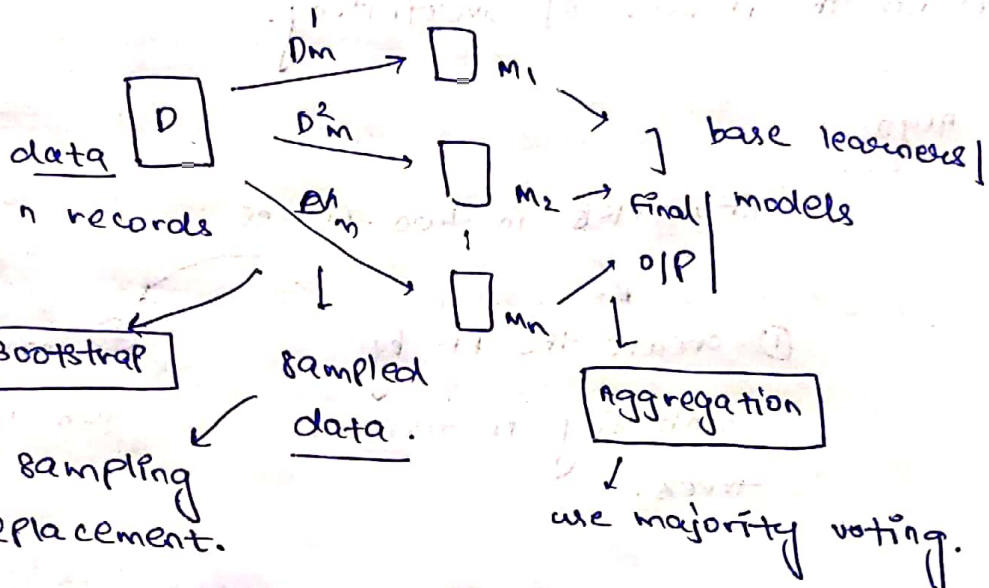
④ repeat ① and ②

⑤ for new data points, find the predictions of each decision tree and assign the category that wins majority votes.

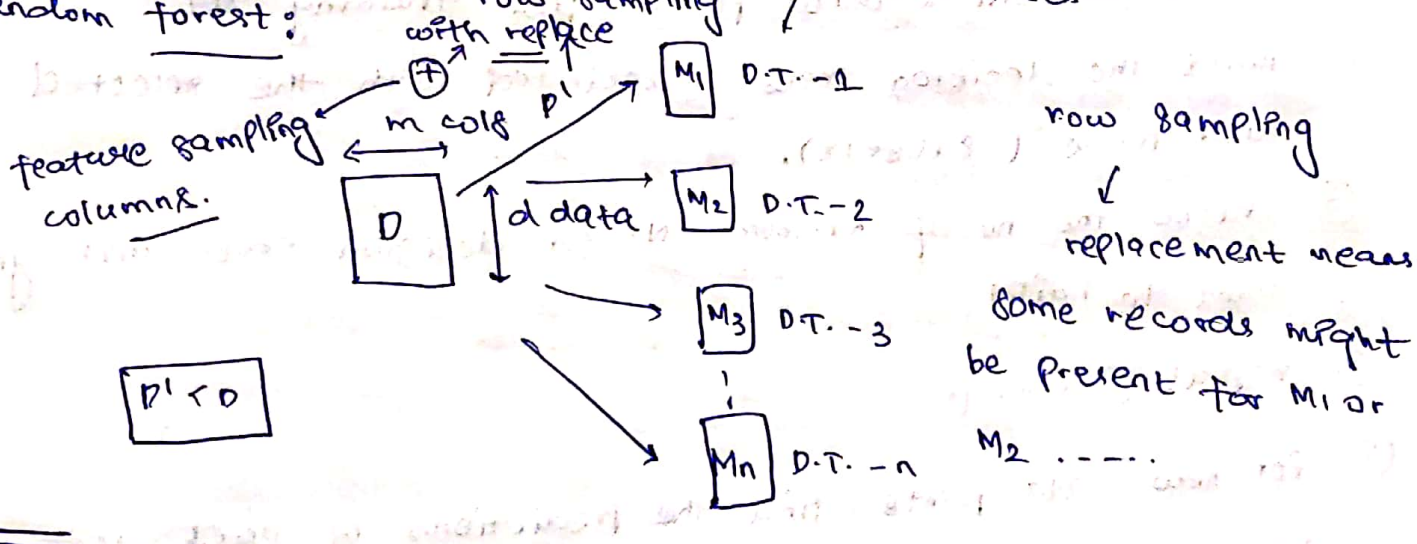
Ensemble Techniques

- Bagging (Bootstrap Aggregation)
  1. Random forest
- Boosting
  1. AdaBoost
  2. Gradient Boosting
  3. XgBoost

Bagging:



Random forest:



Decision Tree

1. Low bias. (training error = less)
2. High variance. (on new data, our model is prone to give larger amount of er

whenever we create a decision tree to its full depth

↳ leads to overfitting.

RF:

① Low bias.

② Low variance. [by taking majority vote]

we do row and col. sampling for every D.T. → if we add new data won't affect RF much.

↓

feature sampling.

in regression → we can take avg. of output of all the decision trees in a RF.

classifier → majority vote.

regressor → mean.
