Random Forest Model

Lecture 15

01/09/2021

Random Forests

- Random Forests are a type of ensemble method.
- An ensemble method is a process in which numerous models are fitted and the results are combined for stronger predictions.
- While this provides great predictions, inference and explainability are often limited.
- Random forests are composed of a number of decision trees where the included predictors are chosen at random.

- One of the most frequently used sampling strategy is the *Bootstrap Aggregating (Bagging).*
- Bagging is a random sampling with replacement.
- A new observation is classified by using all the trees developed in the random forests and majority voting is used for deciding the tree.

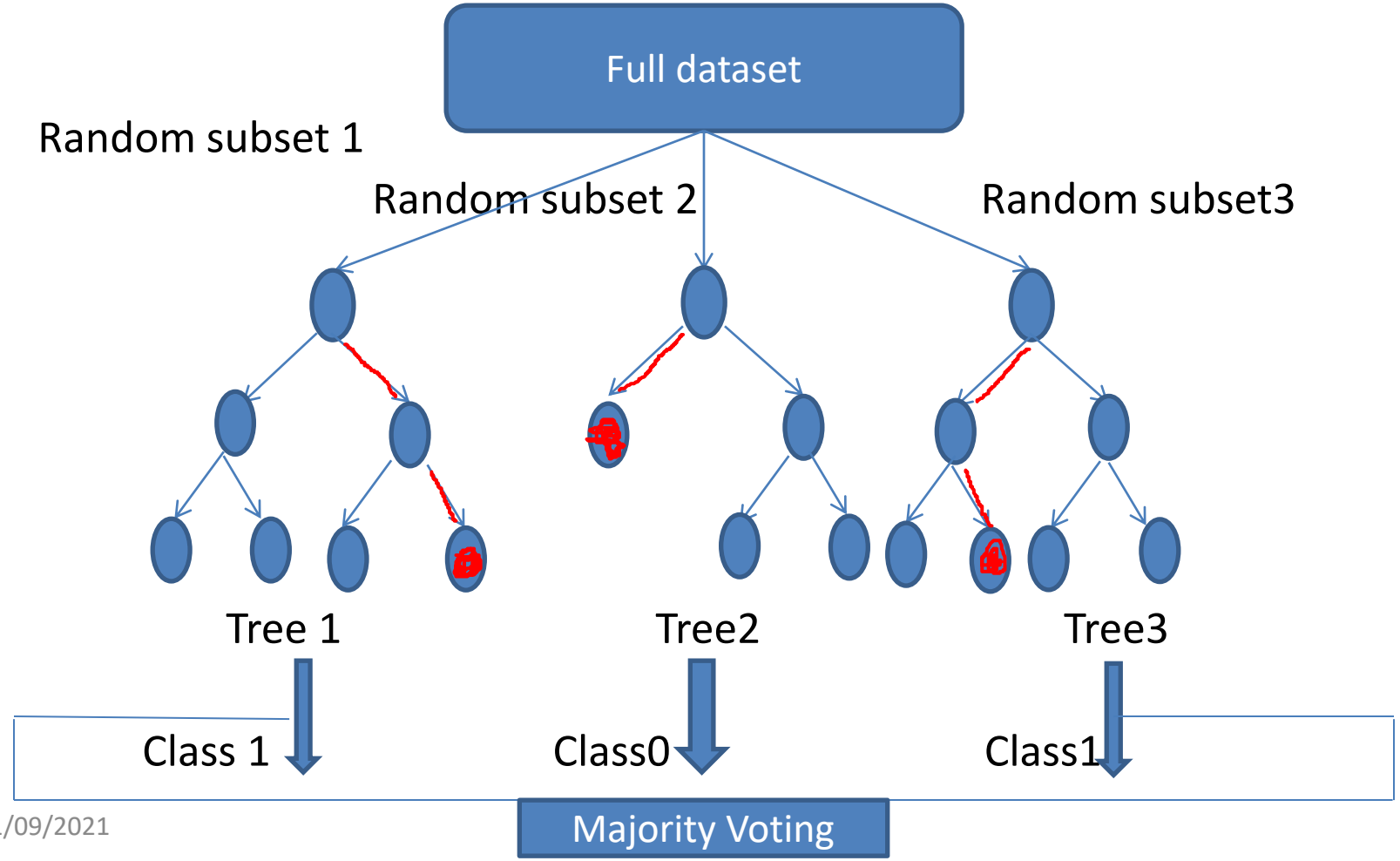
Random forests are developed using the following steps

- 1.Assume that the training data has N observations. One needs to generate several samples of size M (M<N) with replacement (called Bagging). Let the number of samples based on sampling of the training data set be S_1.
- 2.If the data has n predictors, sample m predictors.(m<n)
- 3.Develop trees for each of the samples generated in steps 1 using the sample of predictors from step 2 using CART.
- 4.Repeat step 3 for all the samples generated in step 1.
- 5.Predict the class of a new observation using majority voting based on all trees.

- In general, random forests approach is expected to provide much higher accuracy compared to a single tree.
- However one has to be aware of possible overfitting while using random forests.
- In random forest classifier, if the number of trees is assumed to be excessively large, the model may get over fitted.
- In an extreme case of overfitting, the model may mimic the training data and training error might be almost 0. However, when the model is run on an unseen sample, it may result in a very high validation error.

Random Forest

- Random subset 1



Out-of-bag (OOB) Error

- In random forest, each tree is constructed using a different bootstrap sample from the original data. The samples left out of bootstrap are not used in the construction of i^{th} tree can be used to measure the performance of the model.
- At the end of the run, predictions for each such sample evaluated each time are tallied, and final prediction for that sample is obtained by taking a vote.
- The total error rate of predictions for such samples is termed as out-of-bag(OOB)error rate.

Variance Importance Plot

- It will give you the prediction power of your Random Forest Model. If you drop the top variable from your model, it's prediction power will greatly reduce. On the other hand if you reduce one of the bottom variables, there might not be much impact on prediction power of the model.

Variance Importance Plot