Welcome to Data Science Online Bootcamp

Week#4_Day#3
Prep Material

dφ

Democratizing Data Science Learning

Learning Objectives

Random Forest

Ensembling Models

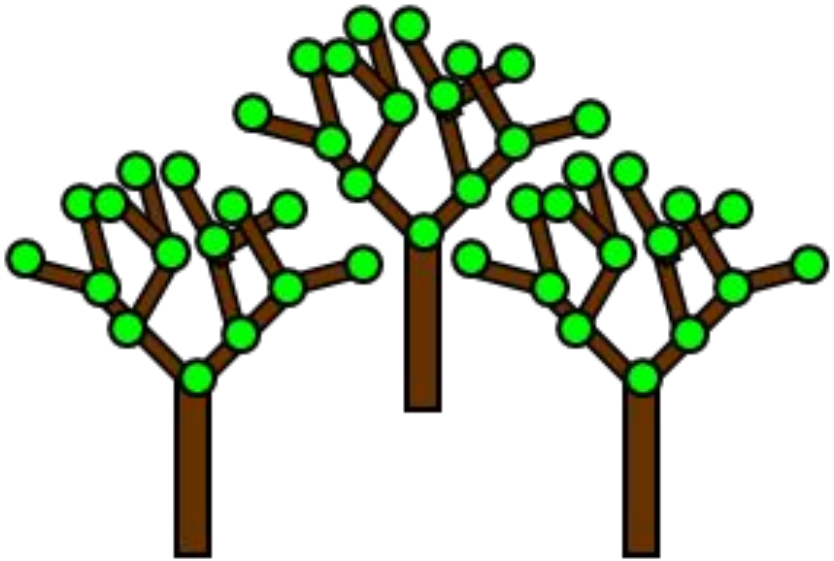
Session Details

Random Forest

Random forest is a flexible, easy to use machine learning algorithm that produces, a great result most of the times even without hyper-parameter tuning.

It is also one of the most used algorithms, because of its simplicity and diversity (it **can be used for both classification and regression tasks**).

Random forest **builds multiple decision trees and merges them together** to get a more accurate and stable prediction.



Applications

- The random forest algorithm is used in a lot of different fields, like banking, the stock market, medicine and e-commerce.

- In finance, for example, it is used to detect customers more likely to repay their debt on time, or use a bank's services more frequently. In this domain it is also used to detect fraudsters out to scam the bank. In trading, the algorithm can be used to determine a stock's future behavior.

- In the healthcare domain it is used to identify the correct combination of components in medicine and to analyze a patient’s medical history to identify diseases.

- Random forest is used in e-commerce to determine whether a customer will actually like the product or not.

Real-life analogy

Andrew wants to decide where to go during one-year vacation, so he asks the people who know him best for suggestions. The first friend he seeks out asks him about the likes and dislikes of his past travels. Based on the answers, he will give Andrew some advice.

This is a typical decision tree algorithm approach. Andrew's friend created rules to guide his decision about what he should recommend, by using Andrew's answers.

Afterwards, Andrew starts asking more and more of his friends to advise him and they again ask him different questions they can use to derive some recommendations for him. Finally, Andrew chooses the places that are recommend the most to him, which is the typical random forest algorithm approach.

Ensemble Models - “The wisdom of crowds”

Let’s pause and think what Andrew did. He took multiple opinions from a large enough bunch of people and then made an informed decision based on them. This is what Ensemble methods also do.

You might have two models that each are good at predicting a certain (different) portion of your dataset. Combining the 2 models into 1 seems like a good idea to increase performance

“ensemble” = Combination of models

Ensemble models in machine learning **combine the decisions from multiple models to improve the overall performance.**

Ensemble Models

So basically ensembling/combining two or more algorithms could improve or boost your performance. But there is a logic behind ensembling...you cannot just randomly combine two models and demand an increase in performance....there is a math behind everything.

So let's dive into the several ensembling methods that you can try out.

Simple Ensemble Techniques

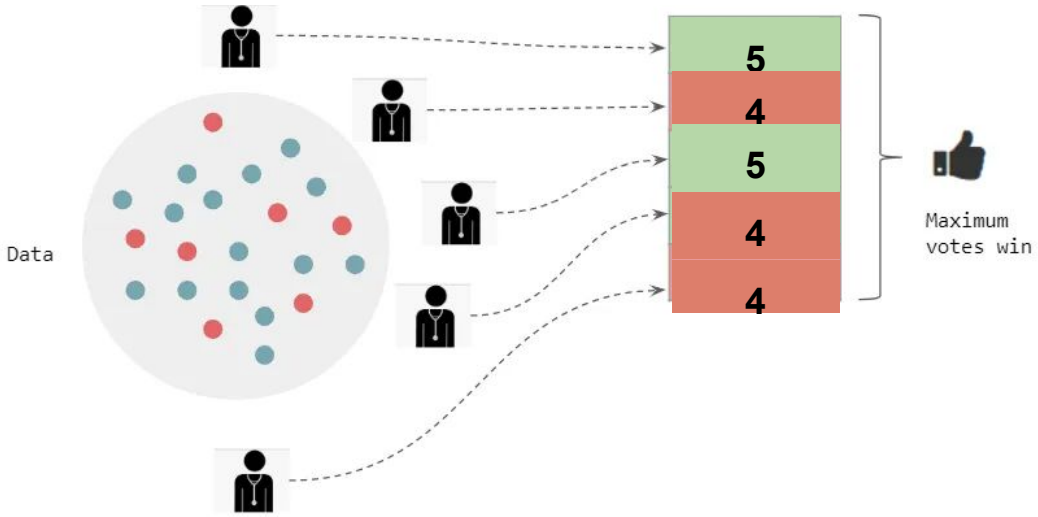
In this section, we will look at a few simple but powerful techniques, namely:

1. Max Voting/ Mode
2. Averaging
3. Weighted Averaging

Max Voting/ Mode

The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. The predictions which we get from the majority of the models are used as the final prediction.

For example, when you asked 5 of your colleagues to rate your movie (out of 5); we'll assume three of them rated 4 while two of them gave a 5. Since the majority gave a rating of 4, the final rating will be taken as 4. You can consider this as taking the mode of all the predictions.



Averaging

In this technique, we take an average of predictions from all the models and use it to make the final prediction.

Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.

For example, in the below case, the averaging method would take the average of all the values.

i.e. (5+4+5+4+4)/5 = 4.4

Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
5	4	5	4	4	4.4

Weighted Average

This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction.

For instance, if two of your colleagues are critics, while others have no prior experience in this field, then the answers by these two friends are given more importance as compared to the other people.

The result is calculated as [(5*0.23) + (4*0.23) + (5*0.18) + (4*0.18) + (4*0.18)] = 4.41.

	Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
weight	0.23	0.23	0.18	0.18	0.18	
rating	5	4	5	4	4	4.41

Advanced Ensemble techniques

Now that we have covered the basic ensemble techniques, we can move on to understanding the advanced techniques, namely:

1. Stacking
2. Blending
3. Bagging Eg: Random forest
4. Boosting Eg: Adaboost, Gradient Boost, Extreme Gradient Boost

We'll learn about Bagging and Boosting techniques. There is nothing to worry about them they are just like any other algorithms like linear, logistic and decision tree.

Bagging (**B**ootstrap **AGG**regat**ING**)

The idea behind bagging is combining the results of multiple models (for instance, all decision trees) to get a generalized result.

Here's a question: If you create all the models on the same training data and combine it, will it be useful? There is a high chance that these models will give the same result since they are getting the same input. So how can we solve this problem? A technique called bootstrapping helps us with that.

Aggregating = Summing or Combining

Bagging combines the different models created by bootstrapping on different sets of training data and hence the name Bootstrap Aggregating.

Random forest is a famous bagging model which uses variations of multiple trees. If same trees are used then it's a bagged decision tree.

Boosting

Here's another question for you: If a data point is incorrectly predicted by the first model, and then the next (probably all models), will combining the predictions provide better results? Such situations are taken care of by boosting.

Intuitively, each new model focuses its efforts on the most difficult observations to fit till now and attempts to correct the errors of the previous model. So at the end of the process, we obtain a strong learner.

Boosting, like bagging, can be used for regression as well as for classification problems.

There are various types of Boosting algorithms which we'll study about soon.

Reading Material

MUST READ

Simple guide for ensemble learning methods: https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2

Slide Download Link

- You can download the slides here: https://docs.google.com/presentation/d/18pyGEZGkEwinktUyZNIhAPAgJMS04tUIIcoyhQFefRw/edit?usp=sharing

Session Details

- **Tutor:** Prasad Seemakurthi
- **Topic:** Ensembling Models and Hyper Parameter Tuning
- **Date & time:** 20th June, at 8:30 pm IST (please locate your time in your timezone here).
- **Youtube live link: https://youtu.be/cQavBseTrQQ**

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack