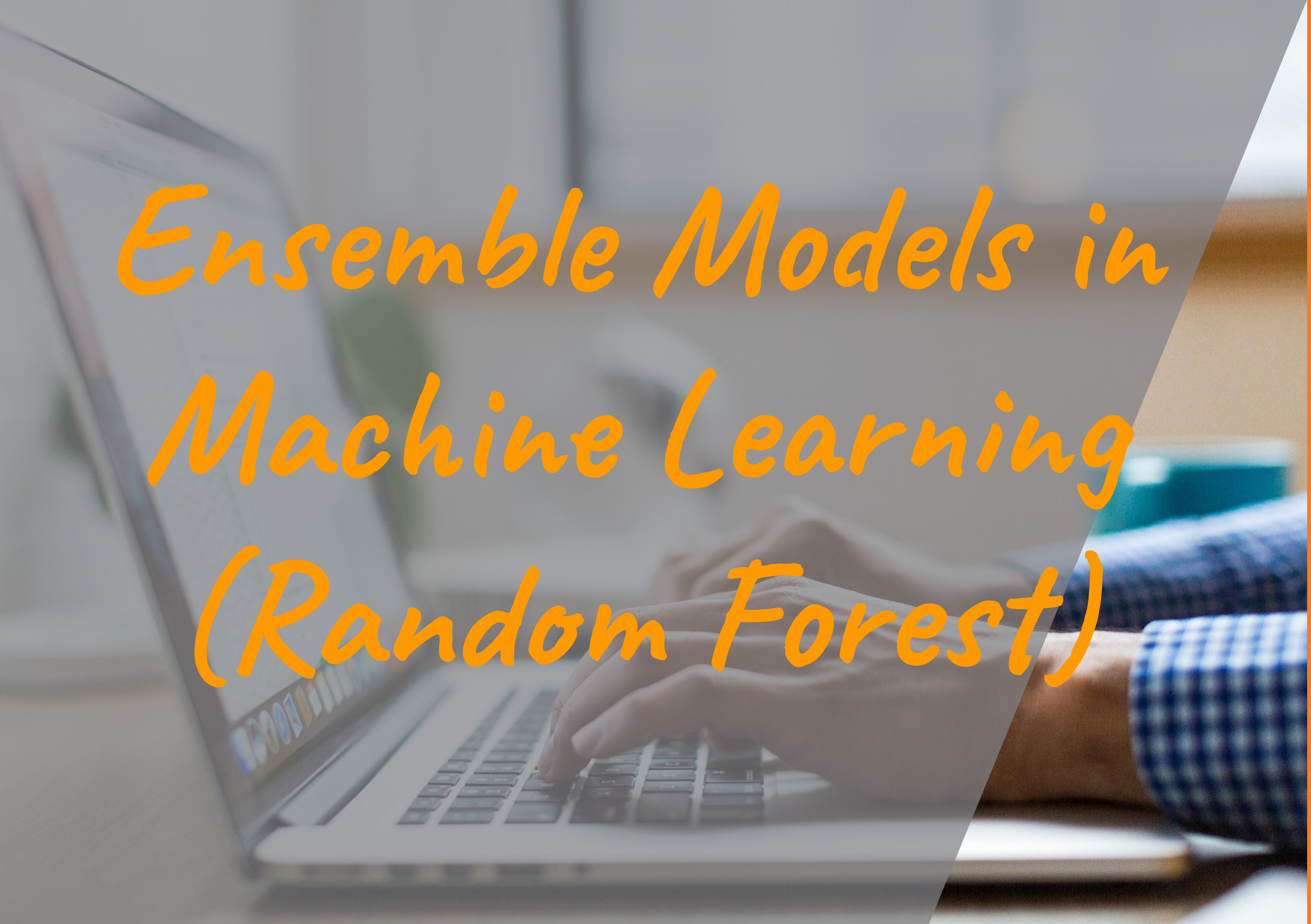

Ensemble Models in Machine Learning (Random Forest)

AGENDA OF THE VIDEO

- What are Ensemble Models?
- Types of Ensemble Models - Bagging and Boosting
- Estimator in Ensemble Models
- Real Life Use Case of Ensemble Models
- Time Constraint in Ensemble Models
- Bagging for Regression and Classification tasks
- Bagging Algorithm (Random Forest)
- Ensemble Models for Random Forest Estimator in Sklearn
- Optimizing our model for best performance
- Task for you.....

What are Ensemble Models?

We have some machine learning algorithms which perform in a certain way. Now, we can improve an individual ML algorithm's performance by doing what is known as "Hyperparameter Optimization" which is a way of tweaking the Hyperparameters of the ML models to come with the best models. Let us not talk about that here...

So, is there any other way of improving a model's performance without doing any sort of "Hyperparameter Optimization". Turns out, there indeed is.

The technique is known as "Ensemble Models". It is just combining many Machine Learning Models to produce powerful results.

Types of Ensembles

1. Understand Various Tree Based Algorithms
    a. Decision Tree
    b. Ensemble of Trees
        i. Bagging Algorithms
            1. Random Forest
        ii. Boosting Algorithms
            1. Adaptive Boosting
            2. Gradient Boosting

A brief discussion on Bagging vs Boosting Ensemble

What is Ensemble?

It is a technique to combine many machine learning models to come with better results than individual models (base estimator).

We split the main_data into subsets of data. How many subsets? Well, that is something, we need to figure out by hit and trial. Then, that many different ML models will be fit on each subset and the to make the prediction, all the models prediction will be combined.

How can we combine many ML models?
It can be done using two different methods. Bagging and Boosting. In Bagging, we create Bags of Models on subset of data and two bags do not have any communication with each other.

In Boosting, we train a ML algorithm on a subset of data and pass its errors to the next subset to boost its performance based on the wrong predictions that the first model made. Hence, the name boosting.

The concept of base_estimator

In Ensemble, we combine many models.

So, what model we combine?

It can be Decision Tree, Logistic Regression or any other Model.

So, this model which we combine is called base_estimator.

What are some real life applications of Ensemble?

1. Mostly ensemble models are used in data science competitions (like on kaggle) to combine many weak learners to produce a powerful learner.
2. Other than that, when some organisations do not have source of sufficient data, then they might use it improve their models performance. But, adding more data if possible is more recommended way of improving the performance of the model that using Ensemble Models.

Time Constraint in Ensemble Learning...

As discussed earlier, In ensemble models, we have decide about the base_estimator and number of base_estimator to combine.

Let us say, base_estimator = Decision Tree and Number of base_estimator = 10

Ensemble Models perform well with large number of base_estimators. This comes at a cost.

Now, training individual Decision Tree model itself is a challenging task, training many of them can get a little costly if you don’t have a powerful computer. With 10 Decision Tree models, we won’t be able to get good results that what we would have got with only one decision tree. To improve the model performance drastically, we need to increase this number which will take some good amount of time.

Bagging for Regression and Classification Tasks

Let us say, we have base_estimator = Decision Tree Classifier
Number of base estimators = 10

Then, Bagging will split the main_data into 10 subsets and on each of them fit the base_estimator.

So, for a classification task, we will get 10 different sets of predictions (class) from each of these different models.

To make the prediction on a new input, it will call every classifier with input and final prediction will be the majority

Bagging for Regression and Classification Tasks (Continued)

Let us say, we have base_estimator = Linear Regression
Number of base estimators = 10

Then, Bagging will split the main_data into 10 subsets and on each of them fit the base_estimator.

Now, for a Regression task, we will get 10 different sets of predictions (continuous value) from each of these different models.

To make the prediction on a new input, it will call every model with input and final prediction will be the average of all 10 different sets of prediction.

Bagging Algorithm (Random Forest)

Given a data set of the form (X,y) where y can either be of categorical type or continuous type,(because Random Forest can be used for both regression and classification tasks).

1. Initialize base_estimator = Decision Tree (For random forest, base estimator is always fixed).
2. Initialize n_base_estimator = number of base estimators
3. Split the data (X,y) in n_base_estimator subsets and fit base_estimator on each of these subset.
4. Finally, for a new input combine the predictions, if classification, use majority voting, or use average of predictions for Regression Related tasks.

Ensemble Models (Bagging or Random Forest) in Sklearn

Optimizing the model's Performance by tweaking few parameters

Task For you..