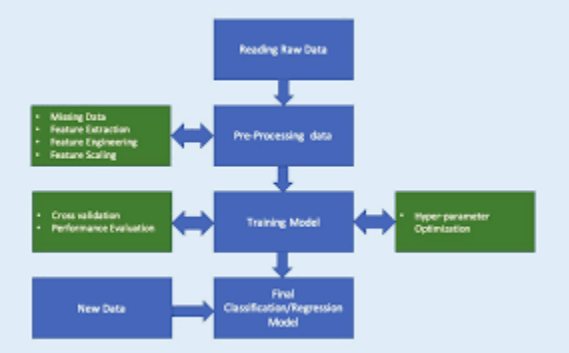
#3 Random Forest

Summary of the 2nd week:



In the second session we focus on new concepts and tools:

- Overfitting vs. Underfitting
- Bias vs. Variance
- Regularization
- Scale the dataset
- cross-validation
- etc

So far, we have had the two most intense weeks. Little by little you will assimilate all this information, and we hope you can see the problems with critical eyes of machine learning.

With critics you should have been with the challenge we put on Saturday. You are free to solve the problems as you see fit. If your intuition tells you to try a logistic regression ... test, tell your partners ... That is the type of thinking we are looking for, so that in the face of projects you look beyond the typical solutions.

You have attached the transparencies and the challenge (with its solution) in module 2.

Objectives 3rd week:

Attentive, we are preparing a very fat one for the 3rd session!

And what concepts are we going to introduce this 3rd week?

- **Hyperparameter tunning**
- **Classification vs Regression Evaluation**
- **Random Forest & Gradient Boosting**
- **Baging, Boosting & Out of Bag**
- **Feature Engineering**
- + review of the above

We will work with videos and notebooks (edited *by us*) from fastAI. In this way, **follow the videos along with the notebooks at the same time**. Do the tests you consider.

Each video lasts an hour and comes with some notes attached. Look carefully at the notes because they are really useful for rereading concepts, code parts, etc. It is not recommended that you watch the video again but go to the notes.

Jump from the video to the notebook, from the notebook to the notes, go back to the video ... Do it in the most comfortable way for you.

Open the notebooks from the colab link if it gives you problems

1º Introduction to Random Forests (from minute 31:00)

https://youtu.be/CzdWqFTmn0Y

- Video notes
- Notebook

This is an introduction video. If it seems too simple, go directly to the notes.

2º Random Forest Deep Dive

https://youtu.be/blyXCk4sqEg

- Video notes
- Notebook (same as above)

3º Performance, Validation and Model Interpretation

https://youtu.be/YSFG_W8JxBo

- Video notes
- Grocery notebook (video from 0 to 50 minutes)
- Random forest interpretation notebook (video from 50 minute)

4º Feature Importance, Tree Interpreter (OPCIONAL)

https://youtu.be/0v93qHDqq_g

- Video notes
- Notebook

In this 4th video, we do not recommend that you stop this week. But if you come back at some point if you are interested. In any case, you have the notes of this 4th video and you can search for the information quickly if you want to take a look at the content.

You already know that any doubt we will be in Slack during the week;) Also remind you that you have a module below to write down project ideas. We will start with the theme of the projects from week 4.

#3 Challenge!

Overview

Good Afternoon! Today we will explore Random Forest - a very powerful way to model data!

A Decision Tree is a very powerful model that can be used alone or as a basis for other models such as Random Forest (RF) and Gradient Boosting (we'll see). In its simplest form, a decision tree asks a series of questions about the characteristics to predict what the result should be. Decision trees also have the added advantage that they can be used for both regression and classification.

A unique decision tree tends to overfitting data. And to solve this problem, **Bagging** is used, a set approach in which N random subsamples of the data set are made using substitution selection and individual decision trees are trained in each subsample. Then the final prediction is the average of all the predictions of the N decision trees.

This is further improved by limiting the characteristic considered in each division to a random subset of characteristics. And it is known as Random Forest ('random forest').

Idea: Random Forest is a model made up of many decision trees that use bootstrapping, random subsets of features and average voting to make predictions.

Resources

Main resources of the session:

- Visual introduction to ML - This is an impressive visualization of how a decision tree works step by step. Take the time to review this and you should have a good fundamental understanding of what is happening under the hood.
- Complete tutorial on decision trees + RF : tutorial that covers the how and why of the use of tree-based models, including decision trees, bagging, RF and Boosting.

An example notebook with which you can solve the challenge

- https://github.com/SaturdaysAI/Itinerario_MachineLearning/blob/master/module_4_decision_trees/4_2_random_forests.ipynb - use the *loans.csv* dataset from previous sessions.

***Optional**: If you prefer we leave other items.*

- https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd
- https://www.datacamp.com/community/tutorials/random-forests-classifier-python

Avanzado: Gradient Tree Boosting

- Guide Gradient Boosting of trees with XGBoost in Python : A complete tutorial using XGBoost for income classification. A good opportunity to review the skills in Python and EDA too!