Introduction to Machine Learning

DataTrek 2021

Présentation

Noel Rignon

B.Ing Genie Logiciel

CEO at FJNR
CTO at ExoFreeMotion

FJNR

Exo
Free
Motion

Github: RignonNoel

LinkedIn: rignonnoel

Summary of the course

Introduction to machine learning

1. Why do we need ML ?
2. What is ML ?
3. ML process
4. Type of ML
5. How to do ML in 2021 ?
6. Exercises
7. Ressources

Why do we need ML ?

We have modern problem

1. Increase in data generation
2. Uncover patterns & trends in data
3. Improve decision making
4. Solve complex problems

We have a lot of data!

Maybe too much sometimes...

In 2020 approximately **1.7MB** of data is created every second for every person on earth

What is ML ?

Arthur Samuel (1959)

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

In short: Ability to learn automatically and improve the performance from a past experience.

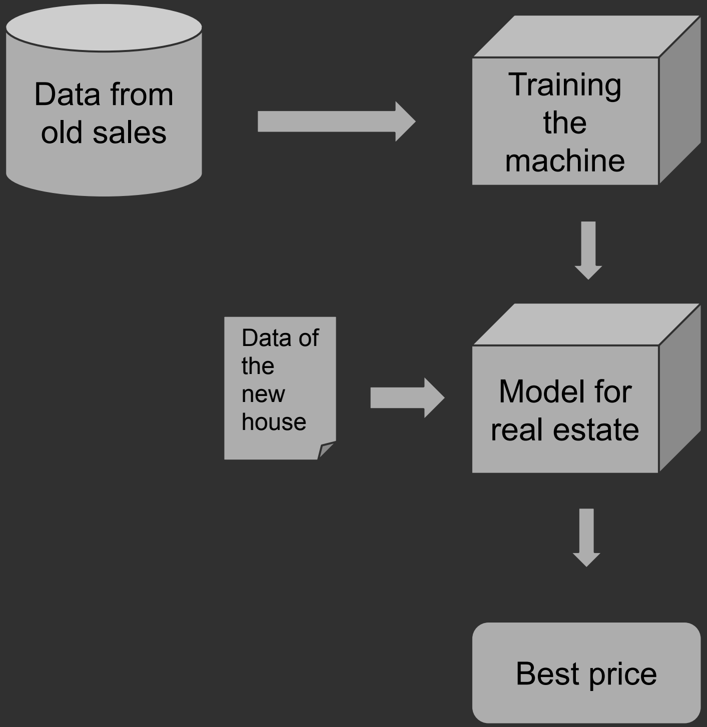
What does it mean ?

If we feed a machine with a lot of data, it can interpret them and “learn” from it to solve more complex problem.

The machine will create a “predictive model” from the data.

Let’s check an example

A real estate agent want to define the best price to sale a house



ML process

With an example

Let’s detail our example

A real estate agent want to define the best price to sale a house

We have a lot of data:

- Price of old sales
- Details of old houses
    - Construction
    - Neighborhood
    - Education / criminality
- Private data about the new house
- Confidential data from some partners

Define the objective

- What do you want to predict ?

  → The price of a specific house

Data gathering

- What kind of data is needed ?
  - → Old sales price
  - → Construction infos
  - → Neighborhood

- Is the data available ?
  - → Yes i already saw peoples using it

- How can i get these data ?
  - → Open data from Montreal
  - → Old intern folder
  - → Info from the customer

Data preparation

- Missing values
  - → Searching for new way to get this data
  - → Remove this data from my database

- Redundant variables
  - → Reformat the file

- Duplicate values
  - → Cleaning

Exploratory data analysis

- Understanding the data
  - What are all the values in the old intern folder ?
  - Searching some theoretical knowledge on the subject

- Correlation between variables
  - Nb of rooms / Price
  - Age / type of construction

- Analysis of pattern
  - If the house is old the price will be better for a little house

Building the model

- Splitting the dataset into two parts (training and testing data)

- Select the good type of algorithm (we will see that just after)

Model evaluation and optimization

- Test the output with the test data

- Measure the accuracy
  - Comparison of output price with real world price signed

- Tuning and improvement of the model

Predictions

- Go play with your new toy!
  → We can change our propositions and estimations based on the final price given by our ML.

Type of ML

Supervised Learning

We train the machine using data which is well labeled.

- You will need labeled data
    - You can buy them
    - You can pay people to label them
    - You can do it yourself
- If the labels are not good the model will be bad

Unsupervised learning

The machine need to figure out the difference by itself and classify the data.

- More complexity
- You will need a test data that is labeled in this case

Reinforcement learning

The machine need to figure out the difference by itself and will learn from “rewards”.

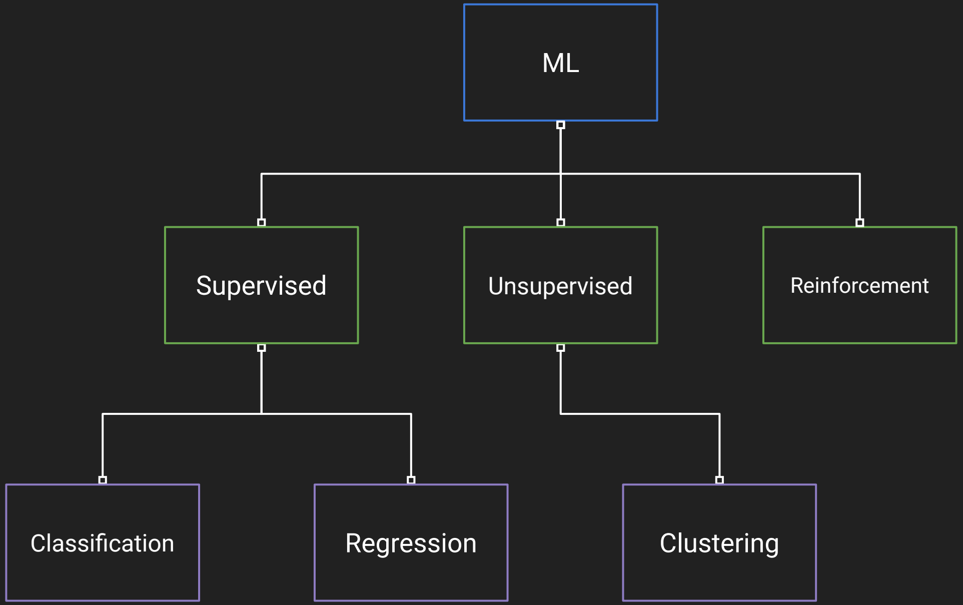
- More advanced subject
- Used only in advanced machine learning areas (ex: self-driving)

Type of problems

Regression: We want to find a continuous value (price, distance, speed)

Classification: We want to find a category (yes/no, authors)

Clustering: We want to defind differents some automatic group and dispatch the inputs in it.



How to do ML in 2021 ?

Do not reinvent the wheel!

Please… do not...

Creating our own ML take a LOT of knowledge to define which kind of algorithm you need and how to do it

You should not underestimate the time and knowledge needed to validate your model and enhance your algorithms

Existing services

There are a lot of services available on the cloud that already created and train an ML for a specific task:

- Facial recognition
- Speech recognition
- Security intrusion

Google Cloud Platform

aws

Azure

ML as a service

Some cloud service can also help you to define your own model

- Contain a lot of algorithm already tested and optimized
- Will put all the algorithm in concurrency to define the best one for your dataset

ML as a service

Google Cloud Platform

Cloud AutoML

- AutoML Vision
- AutoML Video Intelligence
- AutoML Natural Language
- AutoML Translation
- AutoML Tables

ML as a service

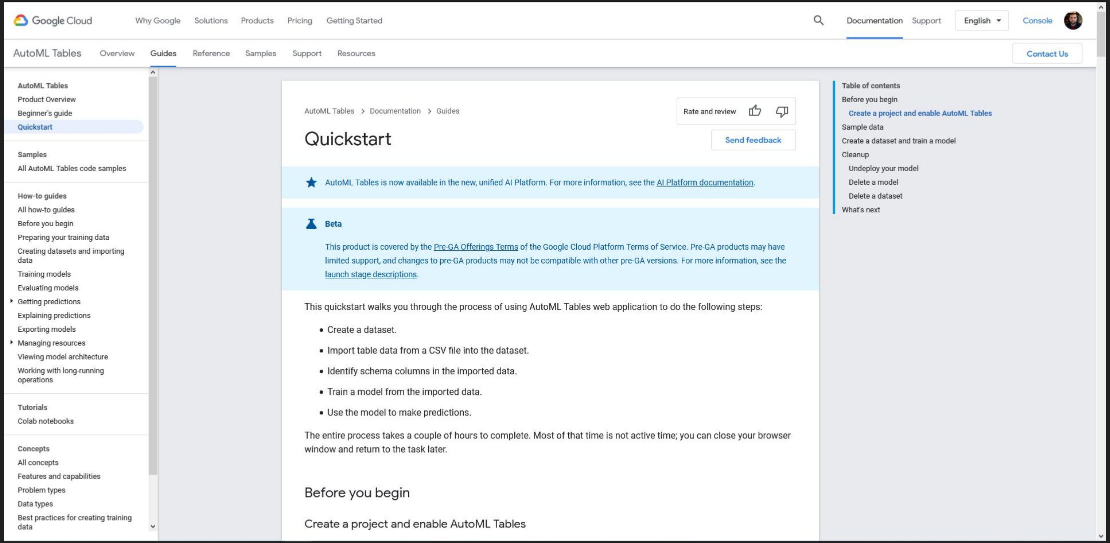
Some cloud provider like Google can provide you some data labeling service if you need some helps

Exercises

Exercises: Introduction to AutoML Tables

Let’s doing the official quick start together!

https://cloud.google.com/automl-tables/docs/quickstart



Quickstart

AutoML Tables is now available in the new, unified AI Platform. For more information, see the AI Platform documentation.

Beta

This product is covered by the Pre-GA Offerings Terms of the Google Cloud Platform Terms of Service. Pre-GA products may have limited support, and changes to pre-GA products may not be compatible with other pre-GA versions. For more information, see the launch stage descriptions.

This quickstart walks you through the process of using AutoML Tables web application to do the following steps:

- Create a dataset.
- Import table data from a CSV file into the dataset.
- Identify schema columns in the imported data.
- Train a model from the imported data.
- Use the model to make predictions.

The entire process takes a couple of hours to complete. Most of that time is not active time; you can close your browser window and return to the task later.

Before you begin

Create a project and enable AutoML Tables

Ressources

Ressource: Google Cloud AutoML

A ML engine ready to use:

https://cloud.google.com/automl/

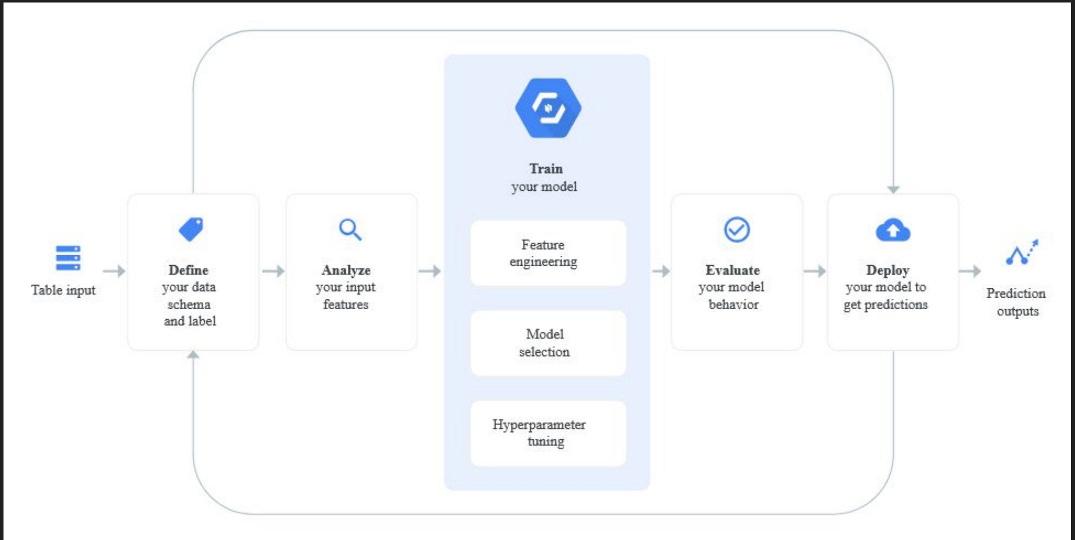


Table input
Define your data schema and label
Analyze your input features
Train your model
Feature engineering
Model selection
Hyperparameter tuning
Evaluate your model behavior
Deploy your model to get predictions
Prediction outputs