**Machine Learning Zoomcamp 2022 Review — Week 01: Introduction to Machine Learning**

I just finished the first week of 'Machine Learning Zoomcamp 2022' by Alexey Grigorev, and to share my notes and progress I will do it through this medium. Weekly articles written as an interview where the main topics learned in the week are addressed, as well as other generalities of the course or recommendations for assignments or projects, I hope you enjoy it, and any feedback will be welcome.

**-To begin with, let’s talk about the main topic of the course, what is “Machine Learning”?**

One way to understand M**achine Learning** is as the process to identify and extract patterns in the data to generate predictions. A combination between attributes of an object (***features***) and information associated with a value to be known (***target***) is the perfect cocktail to create models, which will be used together with different examples of the same attributes, but this time the real values of what is to be predicted will not be considered since this will be the purpose of the model.

**-In the first week the instructor mentions rule-based systems and the difference with machine learning, could you go deeper into the topic?**

Let us suppose that you need a vacation, but you do not want to leave your house alone, so you call your best friend and leave him a handful of chore rules so that your sweet home will be in order when you return. Interestingly, this same approach is used so that a machine can perform a particular task based on an algorithm that delimits its behavior, this type of system is known as **expert systems or rule-based systems**.

Returning to the story, two days later your friend calls to ask you where the cat’s food is, he keeps meowing and you forgot to write it on the list. After a while, an order arrives from a courier company that you never told him you would receive, as a good friend, he calls back to confirm, increasingly unforeseen events occur that you solve after a phone call, but do not let you enjoy your vacation as you wish.

The same thing happens with expert systems, writing all the possible scenarios that a machine could face is impossible, and this is related to the limitations of the hardware that the machine presents, the variability of the data it processes, the complexity of some tasks that due to technical knowledge cannot be included in the algorithm, and for many other reasons.

On the other hand, **Machine Learning** prefers to tell the machine where to go and in turn give it a set of elements that it can use to reach the desired result, of course, the elements could also be a particular set of rules, however, unlike the expert system in which the programmer does most of the work, in machine learning the machine is the one that with the available information will find those patterns that will allow it to reach the expected result.

If we apply this approach to the story, my friend before calling would check the cupboard (which he can open with the key I left him) or call the number of the veterinarian I wrote down in the list and use it to pay the cash I gave him for emergencies and other expenses, this behavior is much closer to the human being, autonomous and intelligent.

**-I have heard that there are many subcategories of Machine Learning, among them Supervised Machine Learning stands out, some words regarding this?**

**Supervised Machine Learning** is teaching a machine to find patterns from examples or training data that are composed of the features (attributes or characteristics of that object) and the target variable (values of the variable to be predicted). As a result, we will obtain a function that later, when applied to other examples, will result in a value close enough to our target value (the task we want to perform).

Using some mathematical notation, the goal of supervised machine learning is to find a function ***g***. When applied to a matrix cap ***X***, values are sufficiently close to our vector ***y*** are obtained.

*g(X) ≈ y*

Where ***g*** is the model, ***X*** is the matrix of features, composed of observations (rows) and features (columns), and, ***y*** is the target variable, with the information we wish to predict.

There are several types of supervised machine learning, and these are differentiated by the value of the target variable.

* If you want to obtain a **numerical value**? We are faced with a ***regression*** problem. An example might be to predict the price in dollars at which you should sell your house.
* If you want to obtain a **category**? We face a ***classification*** problem and this in turn would have other subcategories***, binary classification*** if we have only two categories to classify the result (predict if the stock price will go up or down), and ***multiclass classification*** if we have more than 2 (predict if your favorite soccer team will win, draw, or lose).
* If we want to **sort a set of items** by preferences? We are faced with a ***ranking*** problem, this is widely used by ***recommendation systems***, as your favorite streaming service does by showing you on screen the “Just For You” section, it will assign a high numerical value to those items that match your preferences, and these items will be sorted based on those values.

**-If I have a problem in mind that I want to solve with Machine Learning, is there a methodology or process model that can be used to solve it?**

Yes, one of the most widely used analytical models to organize machine learning projects is the CRISP-DM (Cross Industry Standard Process for Data Mining), which consists of six interrelated and dependent phases, these are:

1) **Business Understanding or Task Understanding** ***(not all problems are business related***): In this phase, it is necessary to define how the problem can be solved and what will be the goal of the project, and what standards will be used to measure it and qualify it as successful. Clarifying if it is necessary to use ML or if another method of less complexity would be sufficient to satisfy our client’s need, is also something to consider during this stage.

2) **Data Understanding:** Here we must analyze the available data sources, decide the requirements, and ask ourselves if the available information is sufficient to achieve the project’s objective. Do not underestimate this phase and be honest with the client, if there are no reliable data sources or large data sets either from internal or external sources, we should rethink the approach to solve the problem, remember that data are an indispensable part of the machine learning model.

3) **Data Preparation:** Transform the data to fit the machine learning algorithm to be used. This phase is characterized by transforming raw data into clean data ready to be used, reformatted, or enriched.

4) **Modeling:** In this phase different machine learning models are trained and the best one is chosen, in this process, it is possible to identify problems with the input data or consider adding new features for a better performance of the model.

5) **Evaluation:** Measure how well the model solves the business problem and if it is good enough to meet the previously agreed standards.

6) **Deployment:** Here we deploy the model to production and make it available to all users, this phase also involves controlling the maintainability and quality of the proposed solution, as well as constantly monitoring user feedback and worrying about scalability, all these tasks are more associated with machine learning engineering.

Along with these steps, there are three additional recommendations I would like to offer:

* **Start simple**, today the computational capacity and the development of machine learning algorithms allow us to process and generate models in a much faster way than a couple of years ago, this allows us to go through the iteration in less time and with fewer costs, which benefits the continuous improvement process by getting feedback in less time from the rest of the phases.
* **It is iterative and interrelated**, a common mistake is to consider each of the phases as a sequential process, and although one phase needs the other, this does not mean that the iteration must be finished to consider a previous phase, do not be afraid if errors are evident in a previous phase, you will always be in time to return and put the cards back on the table, remember that there are associated costs, expectations, time, risk and money, better to be considered and not to arrive in haste to a solution not connected with the problem.
* **Improve**, **communicate, and open channels for feedback**, thanks to the previous two points we understand how easy it is to move between phases, but it is useless if we are not willing to improve in each of them, if something is not clear we must ask as many times as necessary, commonly stakeholders do not even know how to approach a project, They only want a solution without defining its form, that’s where you come in, the data also speak, as well as the models created and the users that use the product, never move away from them, create channels to receive feedback during the whole process and always look for ways to improve, it is impossible to have all the answers.

**-When obtaining the models, what criteria should I use to measure their performance?**

To decide which model to use we could use the **Model Selection Process**, we initially divided our data into two sets: **Training** and **Validation**. We fit our model using the feature matrix and the target variable vector with the **Training Dataset.** This new model will take as input the features matrix of the **Validation Dataset** and generates predictions of the target variable vector, these predictions will be probability values that will try to approximate the real value of the target variable vector for the **Validation Dataset**, then it would only be enough to identify the proportion of predicted values with the real ones, that is, how many values of the total coincide with the expected ones, this is known as **Accuracy**. This series of steps can be repeated but using a different machine learning algorithm each time to obtain different models and compare them, and the accuracy will be the value we will use to choose the model.

The drawback of this process is that machine learning techniques are based on probability so that a good performance of some model may only be due to randomness, and in another data set the results are not as expected, to mitigate this conflict known as **Multiple Comparisons Problem (MCP)** we can add a couple of additional steps to this process, so the stages would be as follows:

1) Divide your data into three different datasets: **Training, Validation, and Test**.

2) With the **Training Dataset** train the models.

3) With the **Validation Dataset** you evaluate them and get the **Accuracy** values.

4) Select the best model based on the best **Accuracy** score or some other performance metric.

5) Apply the selected model to the **Test Dataset**.

6) Compare the performance metrics obtained from both the **Validation** and **Test** **Datasets**, if they match you can be more confident of your correct choice, if not, you may want to consider other options.

**- What prerequisites would you recommend for starting this course?**

After finishing the first week I can say that the content is completely friendly for a beginner level, however, knowing a priori to program in **Python** and being familiar with some of its libraries such as **Pandas** or **NumPy** would make your life easier when solving the task of the week and easily understand the code behind the last three lessons, however, without previous experience with a little more effort and attention you can get a good handful of functions, common and useful operations for Machine Learning. I would also recommend a basic knowledge of **Linear Algebra** and **Probability Theory**, as they also underpin the foundations of these topics and thus understand the mathematical theory underneath the algorithms. Finally, having some background in computer science, cloud computing, containers, and command lines will make it all the way easier,

**-Do you have any recommendations for successfully solving the week’s homework?**

This week’s assignment does not include the theory of the topics seen, it focuses on questions based on data analysis code using the **Numpy** and **Pandas** Python libraries. Therefore, first, it is essential to have your work environment ready and up to date, I recommend installing **Anaconda** in your local, creating a virtual environment, and installing the required libraries, once this is done you will be able to interact with the code.

Second, download the dataset you will interact with, if working with the command line is difficult for you, you can always download the file in your local. What I did was create a folder in my local with the name of the course and organize it by weeks, to have a project that I can upload to GitHub and share with others.

Third, all the knowledge you need to solve the questions is in the last three videos of the week: ***1.7 Introduction to NumPy, 1.8 Linear Algebra Refresher, and 1.9 Introduction to Pandas***. Fourth, read the questions very well and explore the dataset before you start writing code, always have a plan before getting your hands dirty, think about which column may be associated with the type of information you are asked, or try to read between the lines of the question, for example, if it is referring to the most common data of some variable, this could be interpreted as the mode.

Finally, remember to fill out the form correctly so that your assignment is recorded and upload your code to a public site so that it can be reviewed, do not forget to share what you have learned with your peers, this encourages the exchange of ideas and feedback.

**-What did you like most about the Machine Learning Zoomcamp 2022?**

For this I will make a list of the most outstanding ones:

1. The videos are the perfect length, neither so long as to extend the topic with little relevant information, nor so short that the subject is not clear enough.

2. The programming code of the different lessons is useful and easily replicable, thanks to this we can solve future problems in the industry and shield ourselves with the most used functions and algorithms, without the need to overload us with many lines of code that possibly we will not use again.

3. The channels to interact with the community (Slack and Telegram), this program has managed to build an extensive international community of people passionate about data and artificial intelligence, so you will never lack events to attend, books to read, resources to explore, courses to take or people to interact with, and all this aligned to your interests.

4. The dedication and commitment of the instructor Alexey Grigorev for having built this entire Syllabus in an impeccable way and with relevant topics, making it freely available to the community, including new material and redoing the assignments for this new cohort 2022, actively interacting with the students of the course and scheduling weekly sessions to interact with him, simply a master, thank you, for so much.

5. That public learning is promoted, this is something new to me, but I have discovered so many advantages, not only as a way to study to remember more easily all the topics I am covering, but also for what is behind that, to promote the democratization of education and understand that the world works as a whole, and united more is achieved, to start sharing what you do may reach someone who shares your interests and believe me, no doubt, much you can get out of there.

-And finally, how can I get access to Machine Learning Zoomcamp 2022?

All related information can be found at the following link:

[**mlbookcamp-code/course-zoomcamp at master · alexeygrigorev/mlbookcamp-code**  
*We start the course again in September 2022 You can take the course at your own pace. All the materials are freely…*github.com](https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp)

on Alexey Grigorev’s social networks:

* Linkedin:

<https://www.linkedin.com/in/agrigorev/>

* Twitter:

<https://twitter.com/Al_Grigor?s=20&t=0GubFiiB8kfJkQPPGqaaCg>