Python → Data Science$^{\beta}$ → Machine learning algorithms → Typical ML pipeline

Theory: Typical ML pipeline

13 minutes 0 / 4 problems solved

Skip this topic Start practicing

308 users solved this topic. Latest completion was about 5 hours ago.

You might wonder about what machine learning specialists do as part of their job. Of course, different projects mean different tasks but there are some very common steps. In this topic, we will try to highlight them so that you could have a better idea of typical machine learning tasks. This will also help you understand what it takes to become a data scientist.

§1. Data collection

Machine learning is impossible without data. When developing a new machine learning algorithm, experts often use publicly available datasets to benchmark your method and compare it to the ones that are already developed. You can find them on the famous UCI repository, as well as on Kaggle, the largest ML competition platform.

If you are working on a specific problem, let's say, in consultancy, your client (for instance, some company) can already provide you with some data they have which is relevant to the problem they want you to solve. The data can come, for instance, as Excel spreadsheets. Alternatively, you may be given access to a database from where you can load all the necessary data using SQL queries.

However, for some tasks, you might need to collect the data yourself. This can be the case, for example, if you are working on a problem or a particular application no one has worked on before. Data collection can include web scraping, which is automatically extracting and parsing the content of certain web pages, and manually labeling the data.

§2. Data preprocessing

Whether you use available data or collect it yourself, the datasets you end up with can be very messy. Sometimes, there can be a lot of missing values that you might need to fill in somehow. Some values can be simply wrong (imagine someone made a typo when filling in a spreadsheet and inserted 100 instead of 10.0). It is also quite common that the data is coming from different sources. In this case, it is likely that the format is different (for instance, different measure units, date formats, currencies, and so on, used in different files).

So typically the first step in any machine learning project is the data preprocessing. It includes joining data from different sources, dealing with missing values, and so on.

§3. Exploratory data analysis

Once the data is ready to use, it would be a good idea to take a closer look at it before starting the actual modeling part. This step is generally called exploratory data analysis (EDA). Usually, it involves making some plots and calculating some basic statistics on your data.

EDA is a crucial step as it helps you get to know your data better and identify possible problems with it that might have been left unnoticed at the preprocessing step. Besides, at this step, you gain more insights about the data and the events you will be trying to model. This is the time to test some assumptions about the data that you might have and get some ideas on which approaches can be the best to tackle the problem.

§4. Model selection

Now that you know your data well, you can finally start the modeling part! This is typically an iterative process — you start with training an ML model that you believe will do well on the task you are trying to solve. Then, you

Current topic:

Typical ML pipeline ···

Topic depends on:

✕ Intro to Machine Learning ···

Topic is required for:

Introduction to classification ···

Introduction to clustering ···

Introduction to regression ···

Introduction to pandas ···

Introduction to sklearn ···

Table of contents:

↑ Typical ML pipeline
§1. Data collection
§2. Data preprocessing
§3. Exploratory data analysis
§4. Model selection
§5. Deploying your model
§6. Do I have to know all of this?
§7. Conclusions
Feedback & Comments

evaluate the model's performance and carefully investigate it, whether the model performs as expected, whether it has any pattern in the mistakes it makes, and so on. If so, it is also a good idea to devise a method to fix it.

After that, you may want to make the necessary adjustments, train a new model, analyze its performance, and repeat, until you are happy with the model you have.

§5. Deploying your model

Even if you built the greatest ML model in the world, it is of little use if it cannot be used by anyone else than you, or if the results you are getting cannot be reproduced.

To put it simply, at this final stage you need to make sure that your code can be run on any machine, that your implementation is robust (that is, it does not produce unexpected errors), efficient, and scales well.

The process of making your models available in production environments is called **deployment**. For example, a company you are working for can be interested, for instance, in integrating your ML solution into the software they are already using, so you will need to deploy your model so that it can provide predictions to other software systems.

It is typically Machine learning engineers who implement the built model into the production, but in a smaller company, you can be responsible for both developing and deploying ML solutions.

§6. Do I have to know all of this?

Above, we have described the most typical steps in a data scientist's job. It seems like you need to know a lot of things, right?

In some companies (typically, smaller ones), you might be expected to perform all of these tasks by yourself. In others, the roles can be spread across different people; data engineers responsible for data preprocessing, some do the modeling part, the others implement the solution efficiently and deploy it.

You might have noticed that not all of the steps mentioned above are immediately related to machine learning itself but rather to data or software engineering. This is true. Also, ironically, many data scientists report that most of their time is spent exactly on such engineering tasks rather than on pure machine learning (see, for instance, this 2020 Datanami survey).

§7. Conclusions

- Machine learning experts typically deal with very diverse tasks at their job.
- Every ML project is different, but the most common steps are data loading and preprocessing, exploratory data analysis, modeling, and deployment.
- Depending on your job, you can be expected to perform all of these tasks, or the task can be divided among the team.

Report a typo

26 users liked this theory. 0 didn't like it. **What about you?**

😍 🙂 😐 🙁 😡

Start practicing

Comments (0) Hints (0) Useful links (0) Show discussion