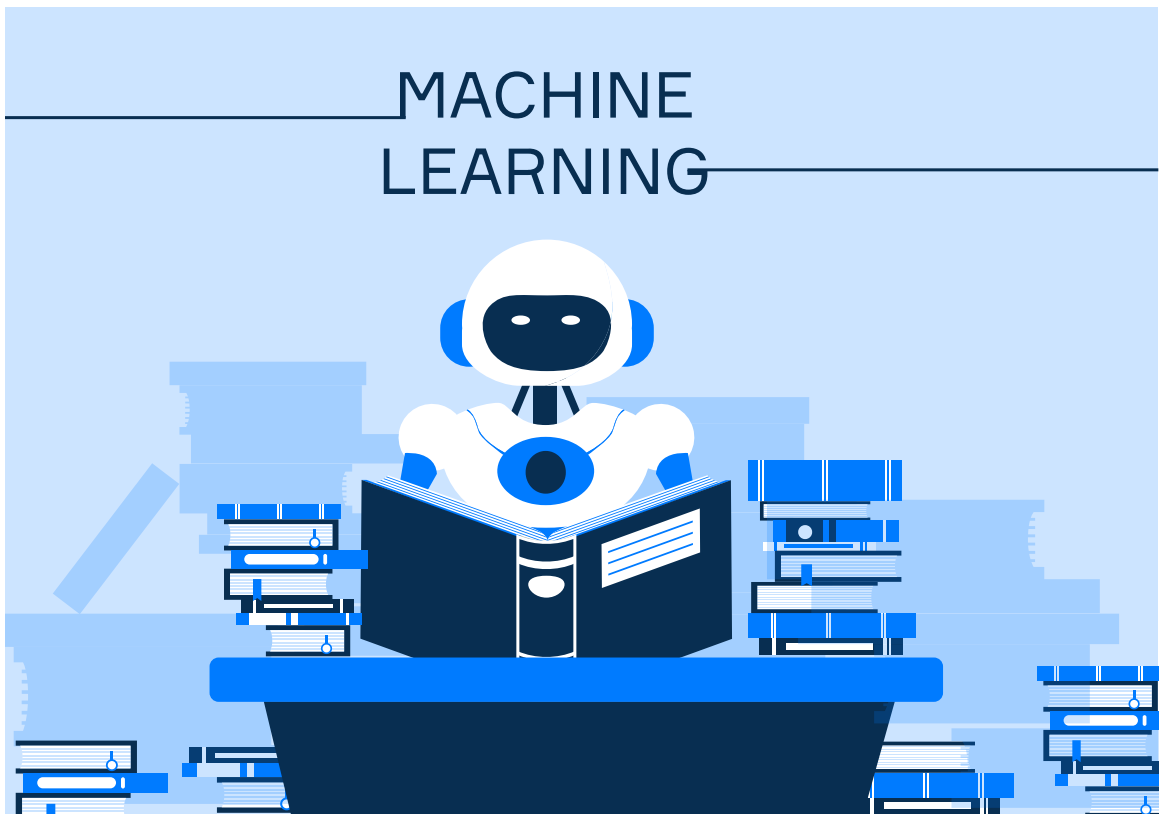
Theory: Intro to Machine Learning

18 minutes 0 / 5 problems solved

When you open your mailbox, you do not see that much of spam these days as they are automatically filtered in a separate folder. If you receive a document in an unknown language, you can easily translate it to any other language you understand in just one click. Your bank blocks your credit card based on some suspicious activity before you even realize that it was stolen. Your favorite streaming service always suggests the movies you fond of, and the special offers you get from your favorite stores are always on point.

What do all of these, and many other technologies share in common besides making our lives way easier? Right, all of them are powered by Machine Learning (ML), a subfield of Artificial Intelligence (AI), which is booming nowadays.

This topic will introduce you to the world of machine learning.



§1. What is machine learning?

You might have heard the term **machine learning** a lot, but what does it actually mean? Well, the goal of ML is to create algorithms that can learn from past experiences and transfer this knowledge to new, unseen cases. Let's consider an example.

Suppose we want to create an ML-based spam filtering system. Then we will need to collect some emails received in the past, both informative and those marked as spam, and introduce them to the algorithm along with their labels. The algorithm, in turn, will try to learn how to distinguish between the two types of mails. Once the learning process is over, our model will be able to analyze new incoming emails and filter out the spam ones.

The key point here is that we do not teach the algorithm ***how*** the two types of emails differ from each other, we just show some examples from the past and let the algorithm figure it out on its own. Cool, right?

Note that the notion of spam may be different for different users of the same mail client. Indeed, while an email about machine learning summer school can be very interesting to you, someone studying medieval music will probably consider it spam. So if you apply the same ML algorithm to different data sets (for example, sets of emails from you and your friend), you will end up with completely different spam filters.

ML can be applied to solve many different problems. Roughly speaking, there are two main ML settings, namely **supervised** and **unsupervised learning**.

Current topic:

Intro to Machine Learning

Topic depends on:

- ✓ Computer algorithms
- ✓ Introduction to Python 18 (Stage 1)

Topic is required for:

Typical ML pipeline

Table of contents:

§2. Supervised learning

In the **supervised learning**, our goal is to learn to predict some **target** attributes from the values of other attributes, often called **features**.

If the target can take on just a few distinct values, the problem is referred to as **classification**. Spam filtering described above is a typical example of a **binary classification** problem. Each email belongs to either of the two categories: spam or regular. In a more general case, there can be more possible classes, and the problem is then called **multi-class classification**. For example, we might want to train an ML model that recognizes hand-written digits. Then, each image of a digit must be associated with one of the 10 classes, from 0 to 9.

Another example of a classification problem is **multi-label classification**. In this setting, the model is assigning not one but multiple binary labels to each example. A typical example of a multi-label classification is text categorization. There is a large number of pre-defined topics (for instance, politics, economy, sports, culture, hobby, ...), and each text can cover several of them (for instance, politics and economy, hobby and sports, and so on). The task is then to predict the correct topics for every text.

If the target attribute of the model is numerical, the problem is referred to as **regression**. An example of a regression problem would be predicting the yearly income of a person based on their education, occupation, background, and other points, or predicting real-estate prices based on the location and size of the property.

§3. Unsupervised learning

Another machine learning setting is **unsupervised learning**. The input data contains no information about the property that we want to predict.

A typical example of an unsupervised ML algorithm is **clustering**. Its goal is to group examples from the training data into so-called clusters, or groups, based on how similar they are. Clustering is often used in market research in order to identify similar groups of consumers based on their purchasing behavior. In bioinformatics, clustering helps categorize genes with similar functionalities and gain insight into structures inherent to populations.

Unsupervised techniques are also commonly used to solve the so-called **anomaly detection** task, the goal of which is to automatically detect suspicious events that are significantly different from the rest of the data. Anomaly detection techniques are widely used by banks to detect fraudulent transactions, in aviation, health monitoring systems, and so on.

§4. Conclusions

- Machine learning algorithms are able to learn from collected data and apply the knowledge to the unknown cases.
- Supervised learning refers to the setting where we want to learn to predict the value of a specific target variable from the data.
- In unsupervised learning, there is no target attribute to predict. The goal is to explore the hidden structure of the data.

Report a typo

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

Start practicing

Comments (4) Hints (0) Useful links (0) Show discussion