Introduction to Machine Learning

Walid Hadri
walid.hadri@student-cs.fr

February 16, 2021

Contents

1 Introduction

The purpose of this lecture is to have a clear idea about what is Machine Learning, what are the subsets of Machine Learning, what is Machine Learning used for, what makes Machine Learning interesting and a hot topic/technology nowadays.

2 What is Machine Learning

While artificial intelligence (AI) is the broad science of mimicking human abilities, machine learning is a specific subset of AI that trains a machine how to learn.

Machine Learning was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum.

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

3 Some applications of Machine Learning

The Machine Learning has so many applications and it has been nearly used in every domain you can think about:

Google Maps: When recommending a route, Google Maps uses all the data collected from other users and companies about the traffic, the speed, the experience... in order to predict the upcoming state of the traffic and recommend the best and fastest route to take.

Social Media: For example behind the Automatic Friend Tagging Suggestions, there is a ML algorithm that learns through the data collected before about you. Facebook uses Face recognition and Image recognition...

Virtual Personal Assistants: Take the example of SIRI or AmazonHome: Speech Recognition, Speech to Text Conversion, Natural Language Processing, Text to Speech Conversion.

Self Driving Cars: Collecting the information with sensors and cameras about the environment using Machine Learning algorithms (Deep Learning).

Dynamic Pricing and risk management: Setting the right price for a good or service is an old problem in economic theory. There are a vast amount of pricing strategies that depend on the objective sought. Be it a movie ticket, a plane ticket or cab fares, everything is dynamically priced. In recent years, artificial intelligence has enabled pricing solutions to track buying trends and determine more competitive product prices. Uber uses the ML algorithm with the nickname Geosurge to fix the price of a ride.

Online Video Streaming and Video Games: they keep collecting information about how you play a video or a game, the time you pause it, you skip things, when you watch, for how long, the rating you give, browsing and scrolling behavior. They use the collected data for purposes like understand the customer behavior, see what they can improve, recommend...

Fraud Detection: one of the most important application is fraud detection. Whenever a customer carries out a transaction – the Machine Learning model thoroughly x-rays their profile searching for suspicious patterns. In Machine Learning, problems like fraud detection are usually framed as classification problems.

4 ML approaches

The approaches will be introduced very briefly in this section, as we will go through them in the following lectures.

4.1 Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. So we have inputs $\mathbf{X}$ and outputs $\mathbf{Y}$, we need to find a mapping function so that $\mathbf{f(X)=Y}$ that has a good performance also on unseen data, when the space of Y is discrete we are dealing with a **classification problem**, otherwise it's a **regression problem** (output is a continuous quantity).

Classification problems like: Mail classification to Spam/NotSpam, Review classification to Positive/Negative, Classification to animal type based on pictures (here we have more than two classes).

Regression problems: like predicting the price of a product, Score prediction, Temperature prediction.

4.2 Unsupervised Learning

Unsupervised Learning is a machine learning technique that learns patterns from untagged data. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data.

The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering, or to determine how the data is distributed in the space, known as density estimation.

One of the most important methods in unsupervised learning is **clustering**: Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. There are other methods like density estimation to understand how the data is distributed and principal component mainly for dimensionality reduction. As you could notice, Unsupervised Learning could be combined with Supervised Learning.

An example to show you how Unsupervised Learning could work: suppose I show you for the first time a dog and a cat, you are going by yourself to identify some characteristics and differences between the two, suppose I show you after another dog and ask you, to which of the two he looks the most. Based on what you have learnt yourself from the first experience, you should answer "it's a dog".

4.3 Semi-supervised Learning

Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training. Semi-supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data), it combines aspects of the former two into a method of its own.

Semi-supervised algorithms operate on data that has a few labels, but is mostly unlabeled. Traditionally, one would either choose the supervised route and operate only on the data with labels, vastly reducing the size of the dataset; otherwise, one would choose the unsupervised route and discard the labels while keeping the rest of the dataset for something like clustering.

This is often the case in real world-data. Since labels are expensive, especially at the magnitude most datasets exist at, large datasets — especially for corporate purposes — may only have a few labels. For instance, consider determining if user activity is fraudulent or not. Out of one million users, the company knows this for ten thousand users, but the other ninety thousand users could either be malicious or benign.

One way to do semi-supervised learning is to combine clustering and classification algorithms. Clustering algorithms are unsupervised machine learning

techniques that group data together based on their similarities. The clustering model will help us find the most relevant samples in our data set. We can then label those and use them to train our supervised machine learning model for the classification task. Say we want to train a machine learning model to classify handwritten digits, but all we have is a large data set of unlabeled images of digits. Annotating every example is out of the question and we want to use semi-supervised learning to create your AI model. So we use some clustering algorithm to construct the clusters (10 clusters in this example, but keep in mind that there are various ways of writing the same digit, so we should pick a greater number of clusters to cover different ways digits are drawn). After training the clustering algorithm, we pick from each cluster the closest element to the centroid that represents better the cluster, and we use the 50 elements we pick as labels to classify after using a classification algorithm from supervised learning. Training the ML algo to classify based on 50 examples sounds like a terrible idea, but bare in mind that in fact, this stated example, which was taken from the book Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, shows that training a regression model on only 50 samples selected by the clustering algorithm results in a 92-percent accuracy. But we can still get more out of our semi-supervised learning system. After we label the representative samples of each cluster, we can propagate the same label to other samples in the same cluster. Using this method, we can annotate thousands of training examples with a few lines of code. This will further improve the performance of our machine learning model.

Another way is Pseudo-labeling, suppose we have a portion of our dataset labeled, we use this labeled data as our training data of the model, we train the model like we are dealing with a supervised problem. Then we use this trained model on the remaining unlabeled data to predict their labels, so they get what we call pseudo-labels. We then train our model on the full dataset, this way we are able to train on a larger dataset without the need of hours of manual labelling. We could be taking a risk of mislabeling data by using pseudo-labeled samples in our training set. Something we could do to lessen the risk is to only include the pseudo-labeled samples in our training set that received a predicted probability for a particular category that was higher than X%. For example, we could make a rule to only include pseudo-labeled samples in the training set that received a prediction for a specific category of, say, 80% or more. This doesn't completely strip out the risk of mislabeling, but it does decrease it. The samples that didn't make the cut due to not having a prediction that met the X% rule could then be predicted on again after the model was retrained with a larger data set that included the first round of pseudo-labeled samples. Also, before going through the pseudo-labeling process, we need to ensure that our model is performing well during training and validation ("well-performing" is subjective here). Additionally, the labeled data that the model was initially trained on should be a decent representation of the full data set.

4.4 Reinforcement Learning

Reinforcement Learning is a subfield of machine learning that teaches an agent how to choose an action from its action space, within a particular environment, in order to maximize rewards over time.It is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. In reinforcement learning, an artificial intelligence faces a game-like situation. The computer employs trial and error to come up with a solution to the problem. To get the machine to do what the programmer wants, the artificial intelligence gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward.

Reinforcement Learning has four essential elements:

- Agent: The program you train, with the aim of doing a job you specify.
- Environment: The world, real or virtual, in which the agent performs actions.
- Action: A move made by the agent, which causes a status change in the environment.
- Rewards: The evaluation of an action, which can be positive or negative.

The main challenge in reinforcement learning lays in preparing the simulation environment, which is highly dependant on the task to be performed. When the model has to go superhuman in Chess, Go or Atari games, preparing the simulation environment is relatively simple. When it comes to building a model capable of driving an autonomous car, building a realistic simulator is crucial before letting the car ride on the street.