✨ Member-only story

Machine Learning Introduction: A Comprehensive Guide

Victor Roman · Follow

Published in Towards Data Science

9 min read · Dec 3, 2018

Listen Share ••• More

Picture from Unsplash

This is the first of a series of articles in which I will describe machine learning concepts, types, algorithms and python implementations.

The main goals of this series are:

1. Creating a comprehesive guide towards machine learning theory and intuition.
2. Sharing and explaining machine learning projects, developed in python, to show in a practical way the concepts and algorithms explained, as well as how they can be applied in real-world problems.
3. Leaving a digital footprint of my knowledge in the subject and inspire others to learn and apply machine learning in their own fields.

The information exposed in this series comes from several sources, being the main ones:

- Machine Learning Engineer NanoDegree (Udacity)
- Python Machine Learning book (by Sebastian Raschka & Vahid Mirjalili)
- Deep Learning with Python book(by Francois Chollet)
- Machine Learning Mastery with Python book(by Jason Brownlee)
- Python Data Science and Machine Learning course by Jose Portilla (Udemy)
- Machine Learning y Data Science con Python course by Manuel Garrido (Udemy)

What Is Machine Learning?

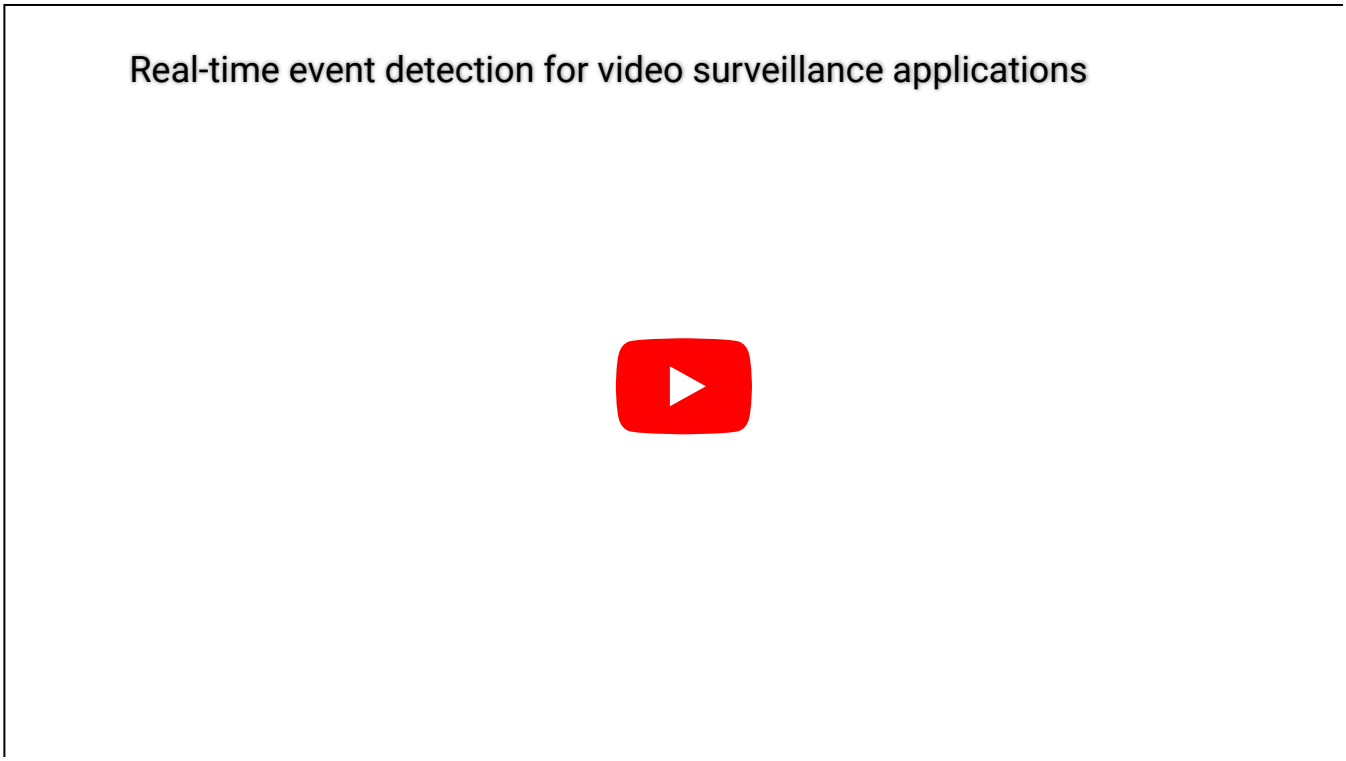
Due to the large decrease in technology and sensors prices, we can now create, store and send more data than ever in history. Up to ninety percent of the data in the world today has been created in the last two years alone. There are 2.5 quintillion bytes of data created each day at our current pace and this pace is only expected to grow. This data feed the machine learning models and it is the main driver of the boom that this science has experienced in recent years.

Machine Learning is one of the subfields of Artificial Inteligence and can be described as:

“Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.” — Dan Fagella

Machine learning offers an efficient way for capturing knowledge in data to gradually improve the performance of predictive models, and make data-driven decisions. It has become an ubiquitous technology and we enjoy its benefits in: e-mail spam filters, self-driving cars, image and voice recognition and world-class go players.

The next video shows a real-time event detection for video surveillance machine learning application.



Basic Terminology and Notations

Generally in machine learning it is used matrix and vector notations to refer to the data. This data is used normally in matrix form where:

- Each separate row of the matrix is a sample, observation or data point.
- Each column is feature (or attribute) of that observation.
- Usually there is one column (or feature), that we will call the target, label or response, and its the value or class that we’re trying to predict.

← Features →					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Table By Author

To train a machine learning model is to provide a machine learning algorithm with training data to learn from it.

Regarding machine learning algorithms, they usually have some inner parameters. ie: In Decision Trees, there are parameters like depth, number of nodes, number of leaves... This inner parameters are called hyperparameters.

Generalization is the ability of the model to make predictions on new data.

Types of machine learning

The types of machine learning that will be studied through this series are:

- Supervised learning
- Unsupervised learning
- Deep learning.

In this series we will explore and study all of the metioned types of machine learning and we will also dig deeper in a kind of deep learning techniques called “reinforcement learning”.

Supervised Learning

Supervised learning refers to a kind of machine learning models that are trained with a set of samples where the desired output signals (or labels) are already known. The models learn from these already known results and make adjustments in their inner parameters to adapt themselves to the input data. Once the model is properly trained, it can make accurate predictions about unseen or future data.

An overview of the general process:

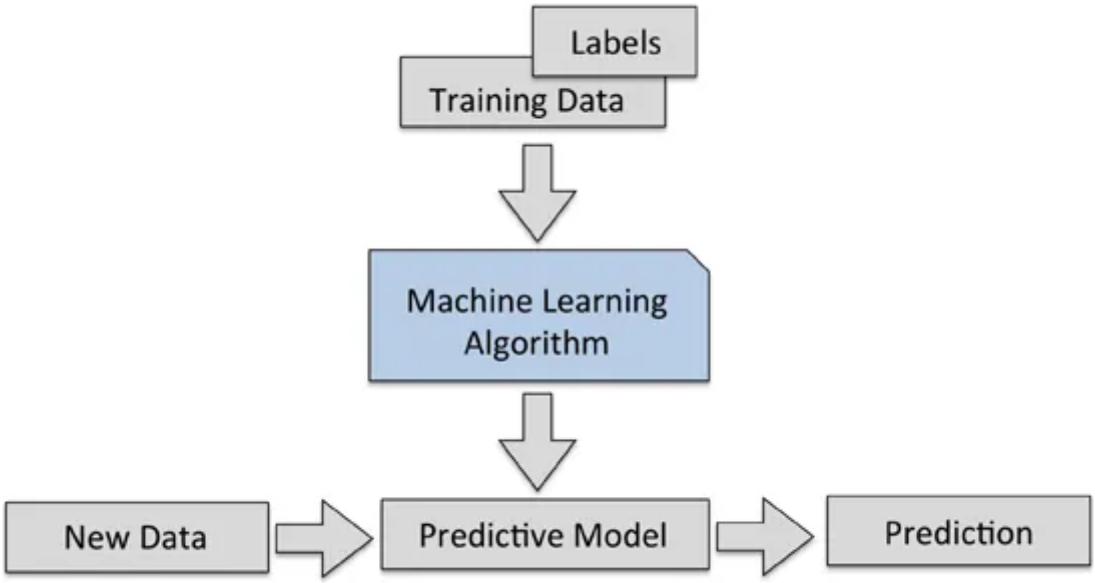


Figure by Author

There are two main applications of supervised learning: classification and regression.

1. **Classsification:**

Classification is a subcategory of supervised learning where the goal is to predict the categorical class labels (discrete, unordered values, group membership) of new instances based on past observations. The typical example is e-mail spam detection, which is a binary classification (either an e-mail is -1- or isn't -0- spam). There is also multi-class classification such as handwritten character recognition (where classes go from 0 to 9).

An example of binary classification: There are 2 classes, circles and crosses, and 2 features, X1 and X2. The model is able to find the relationship between the features of each data point and its class, and to set a boundary line between them, so when

provided with new data, it can estimate the class where it belongs, given its features.

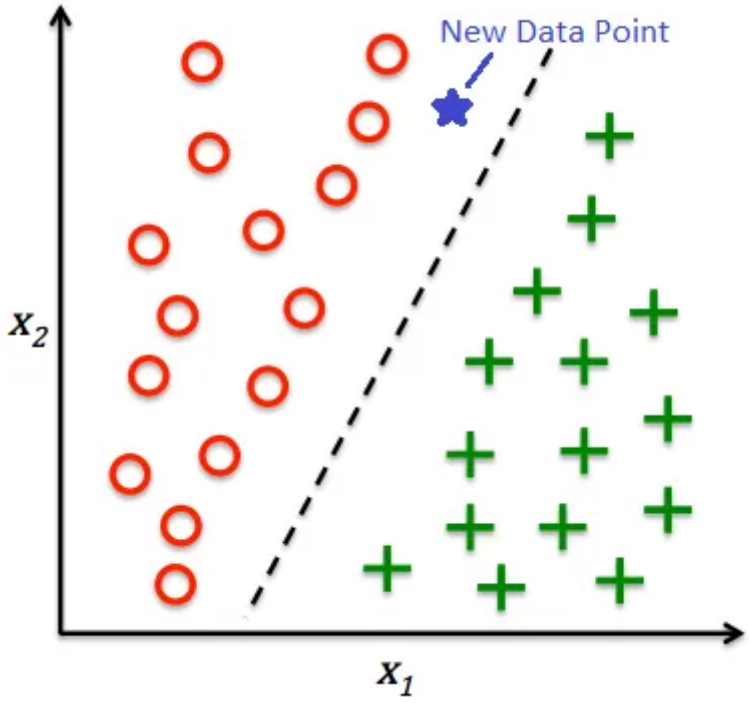


Figure by Author

In this case, the new data point falls into the circle subspace and, therefore, the model will predict its class to be a circle.

2. Regression:

Regression is also used to assign categories to unlabeled data. In this type of learning we are given a number of predictor (explanatory) variables and a continuous response variable (outcome), and we try to find a relationship between those variables that allows us to predict a continuous outcome.

An example of linear regression: given X and Y, we fit a straight line that minimize the distance (with some criteria like average squared distance (SSE)) between the sample points and the fitted line. Then, we'll use the intercept and slope learned, of the fitted line, to predict the outcome of new data.

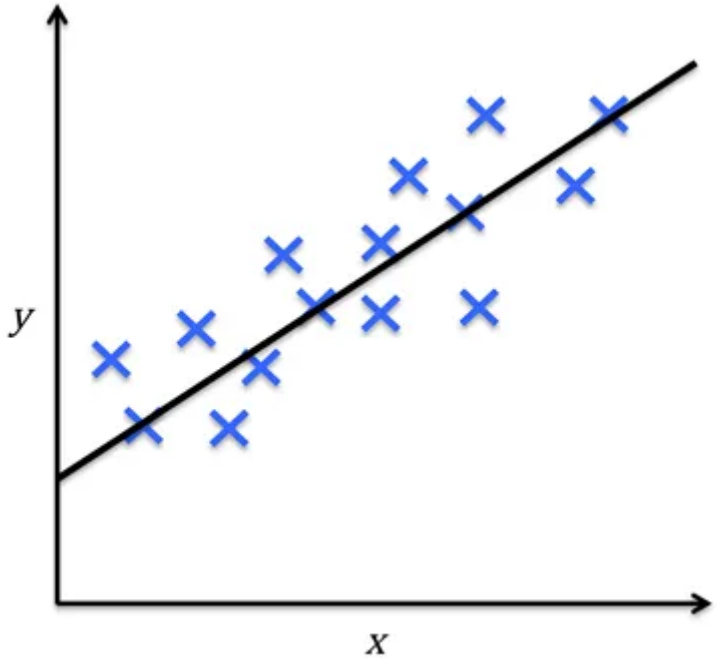


Figure by Author

Unsupervised Learning

In unsupervised learning we deal with unlabeled data of unknown structure and the goal is to explore the structure of the data to extract meaningful information, without the reference of a known outcome variable.

There are two main categories: clustering and dimensionality reduction.

1. **Clustering:**

Clustering is an exploratory data analysis technique used for organizing information into meaningful clusters or subgroups without any prior knowledge of its structure. Each cluster is a group of similar objects that is different to objects of the other clusters.

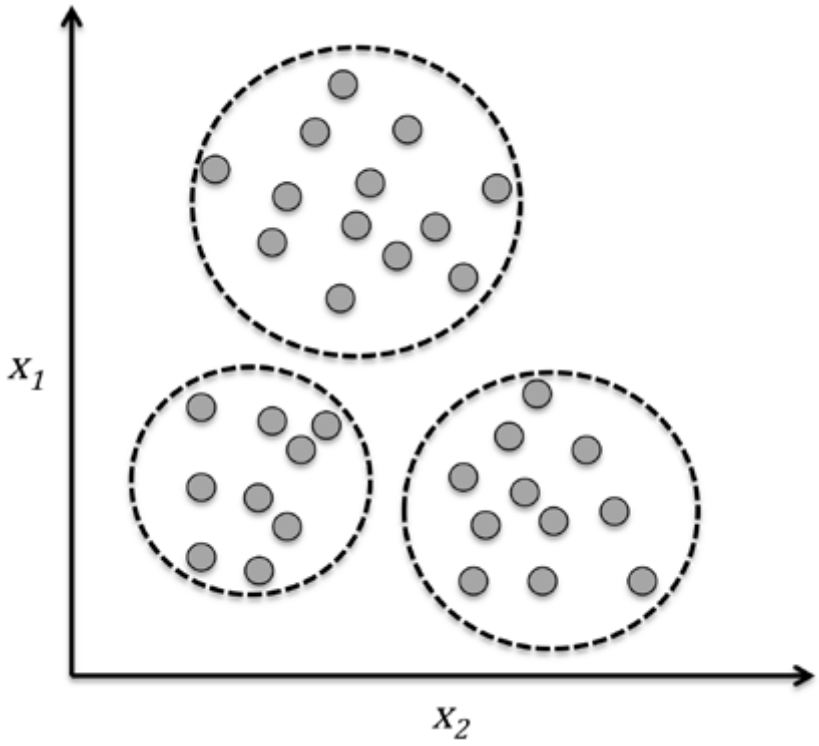


Figure by Author

2. Dimensionality Reduction:

It is common to work with data in which each observation comes with a high number of features, in other words, that have high dimensionality. This can be a challenge for the computational performance of Machine Learning algorithms, so dimensionality reduction is one of the techniques used for dealing with this issue.

Dimensionality reduction methods work by finding correlations between the features, which would mean that there is redundant information, as some feature could be partially explained with the others. It removes noise from data (which can also decrease the model's performance) and compress data to a smaller subspace while retaining most of the relevant information.

Deep Learning

Deep learning is a subfield of machine learning, that uses a hierarchical structure of artificial neural networks, which are built in a similar fashion of a human brain, with the neuron nodes connected as a web. That architechture allows to tackle the data analysis in a non-linear way.

The first layer of the neural network takes raw data as an input, processes it, extracts some information and passes it to the next layer as an output. Each layer

then processes the information given by the previous one and repeats, until data reaches the final layer, which makes a prediction.

This prediction is compared with the known result and then, by a method called backpropagation, the model is able to learn the weights that yield accurate outputs.

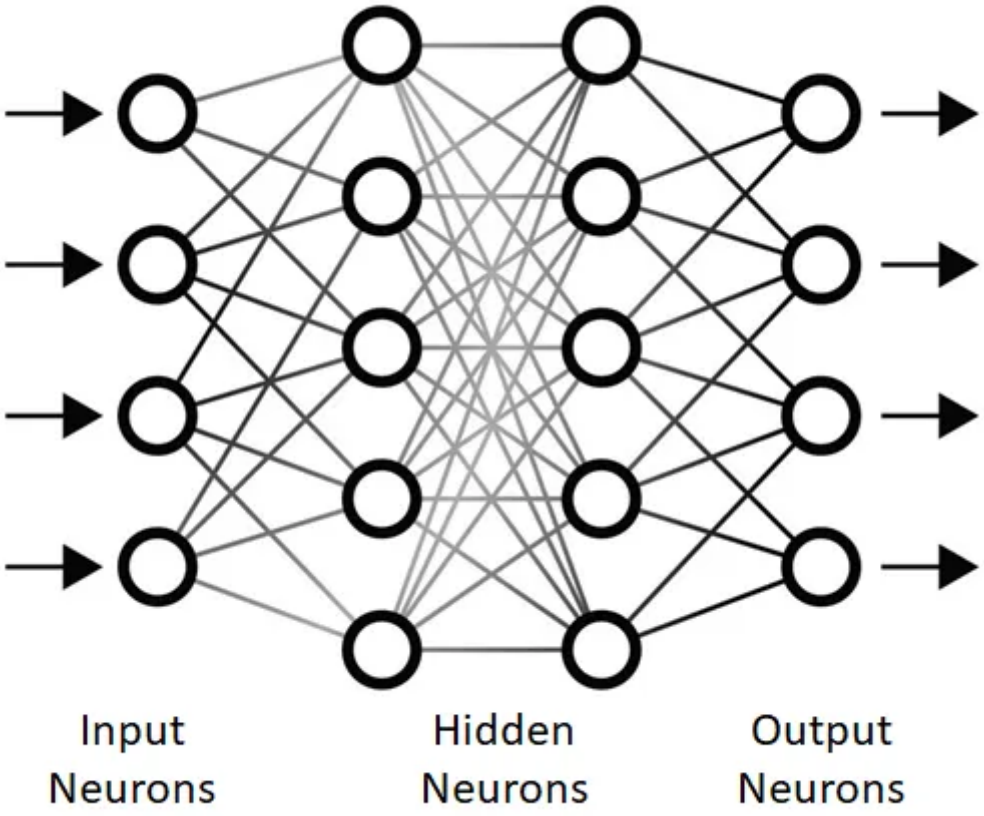


Figure by Author

Reinforcement Learning

Reinforcement learning is one of the most important branches of Deep Learning. The goal is to build a model in which there is an agent that takes actions and where the aim is to improve its performance. This improvement is done by giving an specific reward each time that the agent performs an action that belongs to the set of actions that the developer wants the agent to perform.

The reward is a measurement of how well the action was in order to achieve a predefined goal. The agent then uses this feedback to adjust its future behaviour, with the objective of obtaining the most reward.

One common example is a chess engine, where the agent decides from a series of possible actions, depending on the board's disposition (which is the environment's state) and the reward is given when winning or loosing the game.

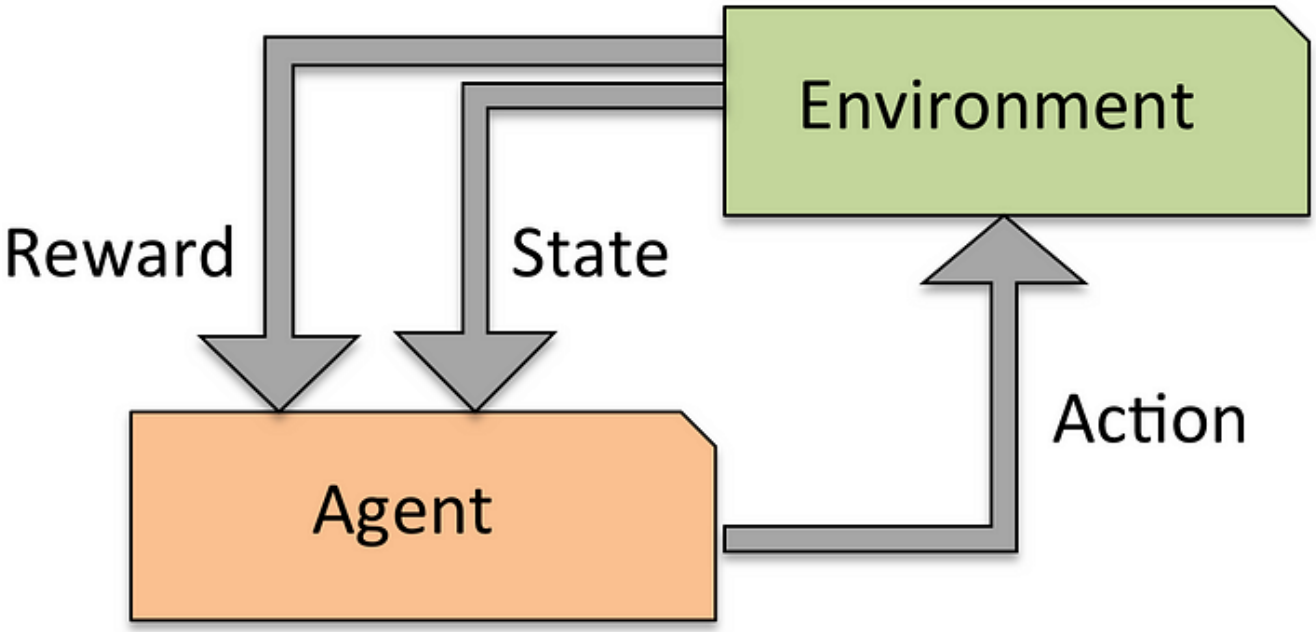


Figure by Author

General Methodology for Building Machine Learning Models

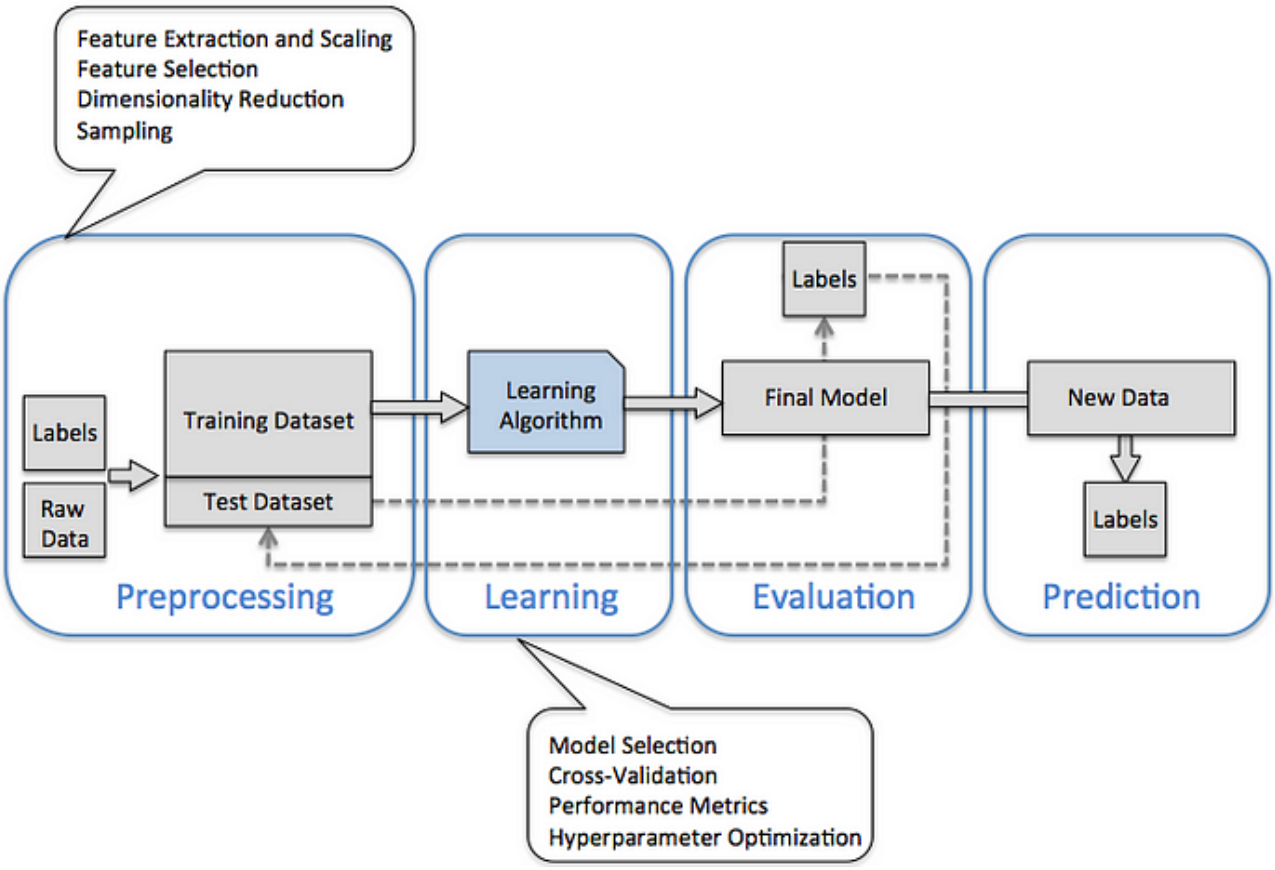


Figure by Author

Preprocessing:

It is one of the most crucial steps in any Machine Learning application. Usually data comes in a format that is not optimal (or even inadequate) for the model to process it. In that cases preprocessing is a mandatory task to do.

Many algorithms require the features to be on the same scale (for example: to be in the [0,1] range) for optimizing its performance, and this is often done by applying normalization or standardization techniques on the data.

We can also find in some cases that the selected features are correlated and therefore, redundant for extracting meaningful information from them. Then we must use dimensionality reduction techniques to compress the features to smaller dimensional subspaces.

Finally, we'll split randomly our original dataset into training and testing subsets.

Training and Selecting a Model

It is essential to compare a bunch of different algorithms in order to train and select the best performing one. To do so, it is necessary to select a metric for measuring the model's performance. One commonly used in classification problems is classification accuracy, which is the proportion of correctly classified instances. In regression problems one of the most popular is Mean Squared Error (MSE), that measures the average squared difference between the estimated values and the real values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points,
f_i the value returned by the model and
y_i the actual value for data point i.

Figure bu Author

Finally, we will use a technique called cross-validation to make sure that our model will perform well on real-world data befiore using the testing subset for the final evaluation of the model.

This technique divides the training dataset into smaller training and validating subsets, then estimates the generalization ability of the model, in other words, estimating how well it can predict outcomes when provided with new data. It then repeats the process, *K* times and computes the average performance of the model by dividing the sum of the metrics obtained between the *K* number of iterations.

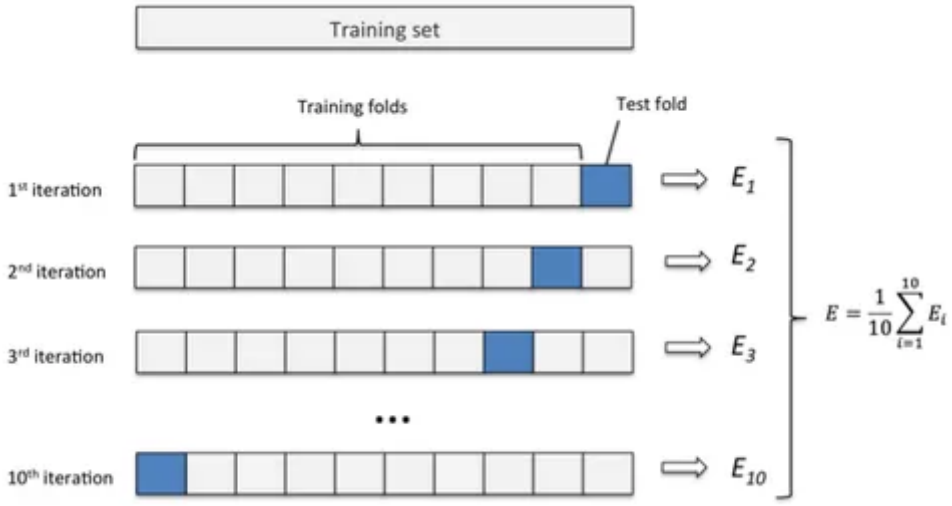


Figure by Author

In general, the default parameters of the machine learning algorithms provided by the libraries are not the best ones to use with our data, so we will use hyperparameter optimization techniques to help us to do the fine tunning of the model's performance.

Evaluating Models and Predicting with New Data

Once we have selected and fitted a model to our training dataset, we can use the testing dataset to estimate the performance on this unseen data, so we can make an estimation of the generalization error of the model. Or evaluate it using some other metric.

If we are satisfied with the value of the metric obtained, we can use then the model to make predictions on future data.

Wrap Up

In this article we learned what is Machine Learning painting a big picture of its nature, motivation and applications.

We also learned some basic notations and terminology and the different kinds of machine learning algorithms:

- Supervised learning, with classification and regression problems.
- Unsupervised learning, with clustering and dimensionality reduction.
- Reinforcement learning, where the agent learn from its environment.
- Deep learning and their artificial neuron networks.

Finally, we made an introduction to the typical methodology for building Machine Learning models and explained its main tasks:

- Preprocessing.
- Training and testing.
- Selecting a model.
- Evaluating.

If you liked this post then you can take a look at my other posts on Data Science and Machine Learning here.

If you want to learn more about Machine Learning, Data Science and Artificial Intelligence ***follow me on Medium****, and stay tuned for my next posts!*

Machine Learning | Deep Learning | AI | Supervised Learning | Unsupervised Learning

Follow

Written by Victor Roman

2.3K Followers · Writer for Towards Data Science

Industrial Engineer and passionate about 4.0 Industry. My goal is to encourage people to learn and explore its technologies and their infinite posibilites.

More from Victor Roman and Towards Data Science

Victor Roman in Towards Data Science

Machine Learning Project: Predicting Boston House Prices With Regression

Learn how to apply regression to solve a real-world problem

· 17 min read · Jan 20, 2019

617 4

Jacob Marks, Ph.D. in Towards Data Science

How I Turned My Company's Docs into a Searchable Database with OpenAI

And how you can do the same with your docs

15 min read · Apr 25

Leonie Monigatti in Towards Data Science

Getting Started with LangChain: A Beginner's Guide to Building LLM-Powered Applications

A LangChain tutorial to build anything with large language models in Python

✦ · 12 min read · Apr 25

2.2K 18

Victor Roman in Towards Data Science

How To Develop a Machine Learning Model From Scratch

In this article we are going to study in depth how the process for developing a machine learning model is done. There will be a lot of...

✦ · 16 min read · Dec 23, 2018

1.4K 4

See all from Victor Roman

See all from Towards Data Science

Recommended from Medium

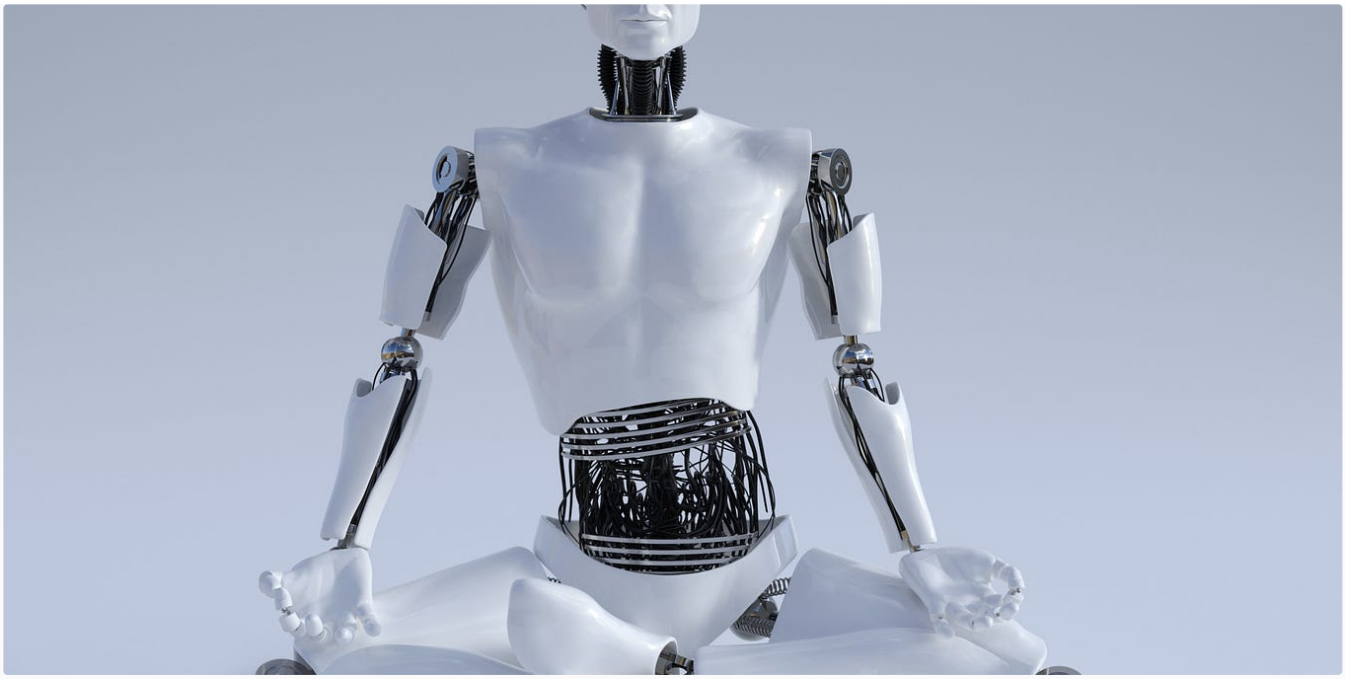
Matt Chapman in Towards Data Science

The Portfolio that Got Me a Data Scientist Job

Spoiler alert: It was surprisingly easy (and free) to make

✨ · 10 min read · Mar 24

2.9K 44



The PyCoach in Artificial Corner

You're Using ChatGPT Wrong! Here's How to Be Ahead of 99% of ChatGPT Users

Master ChatGPT by learning prompt engineering.

· 7 min read · Mar 17

Lists

What is ChatGPT?

9 stories · 64 saves

Staff Picks

329 stories · 83 saves

Visual Storytellers Playlist

26 stories · 12 saves

Alexander Nguyen in Level Up Coding

Why I Keep Failing Candidates During Google Interviews...

They don't meet the bar.

· 4 min read · Apr 13

4.2K 127

Albers Uzila in Level Up Coding

Wanna Break into Data Science in 2023? Think Twice!

It won't be smooth sailing for you

· 11 min read · Dec 23, 2022

874 14

Leonie Monigatti in Towards Data Science

Getting Started with LangChain: A Beginner's Guide to Building LLM-Powered Applications

A LangChain tutorial to build anything with large language models in Python

· 12 min read · Apr 25

2.2K 18

Aleid ter Weel in Better Advice

10 Things To Do In The Evening Instead Of Watching Netflix

Device-free habits to increase your productivity and happiness.

· 5 min read · Feb 15, 2022