Dispensa Intro to Machine Learning

Bonmassar Ivan

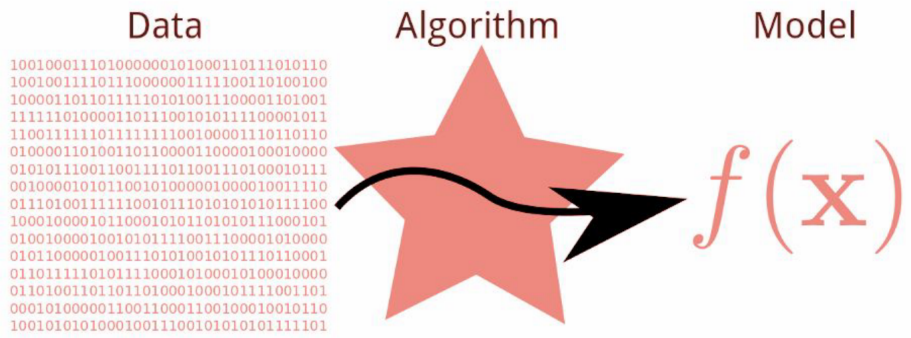
June 28, 2022

Contents

Chapter 1

ML Basics

The main notion to get from this is the following. ML allows computers to gain **knowledge** acquired through **algorithms** by learning from data. This knowledge is represented through a **model** which is then used on future data.



The training data produces a model or predictor, whereas the testing data produce a prediction.

1.1 Data

Data can be a list of movies from IMDB which is easily representable. Although we need to use **features** when taking into consideration other examples. Features is how an algorithm view data and is generally represented with vectors.

For example, classifying different apples, features could be the shape and colour of them.

The general problem with data for classification for example is that not all data is the same. For instance a banana can either be green or yellow. Although this is true we cannot go to deep with this so we use a probabilistic model called **data generating distribution**. Both training and test data are based on this.

So in our previous example, we will generalize and say that bananas are yellow.

Definition 1.1.1 (Probability distribution)**.** Describes how likely certain events are.

High probability: round apples
Low probability: square apples

1.2 Types of Learning

SUPERVISED LEARNING
Supervised learning is when the algorithm is given labeled examples and the predictor should output a label. A further example of this is **classification**, where the model classifies from a pool of categories.

Given a training set $T = (x_i, y_i)$ learn a function that predicts y given x. x is multi-dimensional.

Some real world examples can be facial recognition, spam detection and character recognition.

Regression is similar to classification only with real values (i.e. numbers).

Ranking the label is a ranking (most similar, most popular web pages etc.)

UNSUPERVISED LEARNING The given data is without labels. Some examples are:

Clustering, where the output is the general structure of the data set (clusters of data). Real world examples are image segmentation, social network analysis

Anomaly detection

Dimensionality reduction

REINFORCEMENT LEARNING

The idea is that the agent interacts with the environment and receives rewards based on behavior.

1.3 ML Ingredients

TASK

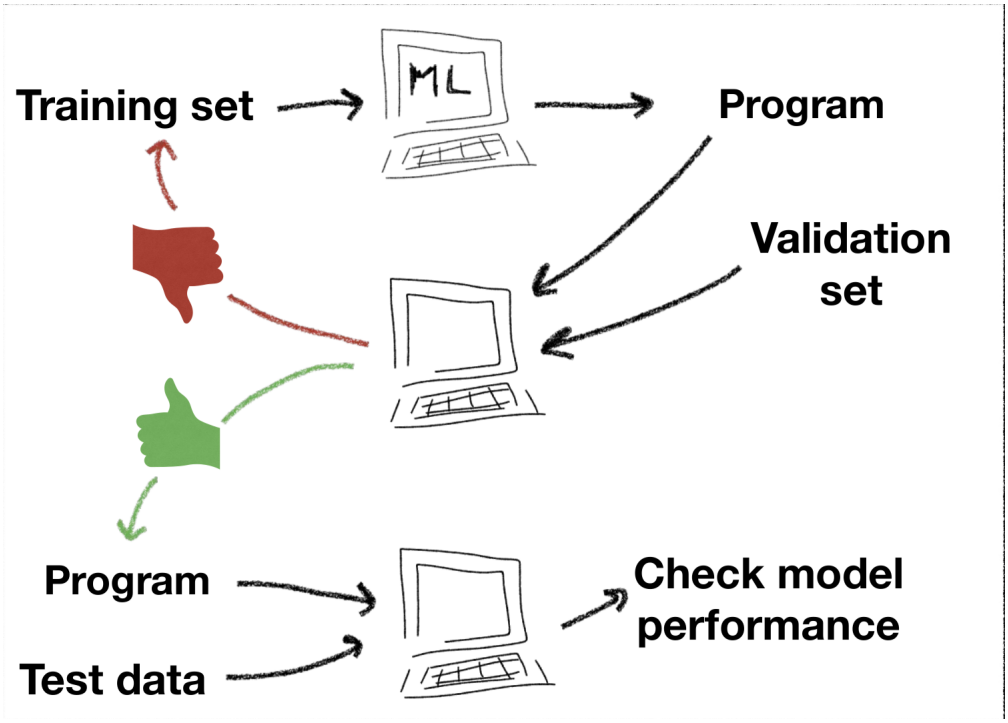
A task represents the type of prediction being made to solve a problem.

Assigning each input $x \in \mathscr{X}$ to an output $y \in \mathscr{Y}$

Data

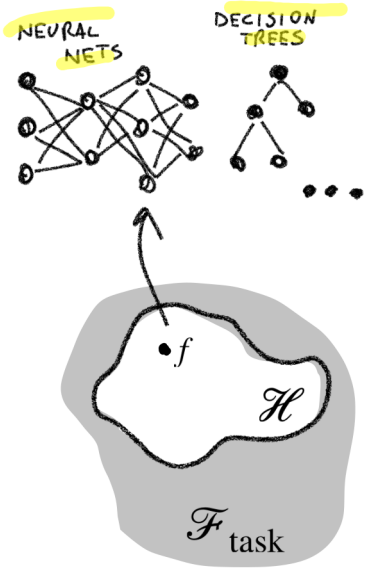
Data is basically the information required to solve a specific problem and as said previously is usually sampled from an unknown data generating distribution :

$\mathbf{p}_{data}$
For classification and regression $\mathbf{p}_{data} \in \triangle(\mathscr{X} \times \mathscr{Y})$



Model and hypothesis space

A model is like a program that solves the problem. There are various models (decision trees, neural networks..) and a set of them makes up the hypothesis space.



The objective

The objective is to minimize an error function $E(f, \mathbf{p}_{data})$ to find the optimal function

$f^\star = \arg\min E(f, \mathbf{p}_{data})$.

This however is really hard to do because of the too large search space.

The feasible target is the optimal one in a restricted hypothesis space $\mathscr{H}$

$f^\star_{\mathscr{H}} = \arg\min E(f, \mathbf{p}_{data})$.

This is also not doable because we do not have access to $\mathbf{p}_{data}$

The **actual target** is then the following:

$f^\star_{\mathscr{H}}(\mathscr{D}_n) = \arg\min E(f, \mathscr{D}_n)$. where $\mathscr{D}_n$ is a training set.