Machine learning, data science, and artificial intelligence

Clément Levallois

2017-31-07

Table of Contents

last modified: 2017-11-04

1. Explaining machine learning in simple terms

a. A comparison with classic statistics

Let's compare machine learning to something we would call "regular statistics":

A basic method in statistics is to compute a regression line to identify a trend from a scatter plot.

To illustrate, we take some data about marketing budgets and sales figures in the corresponding period:

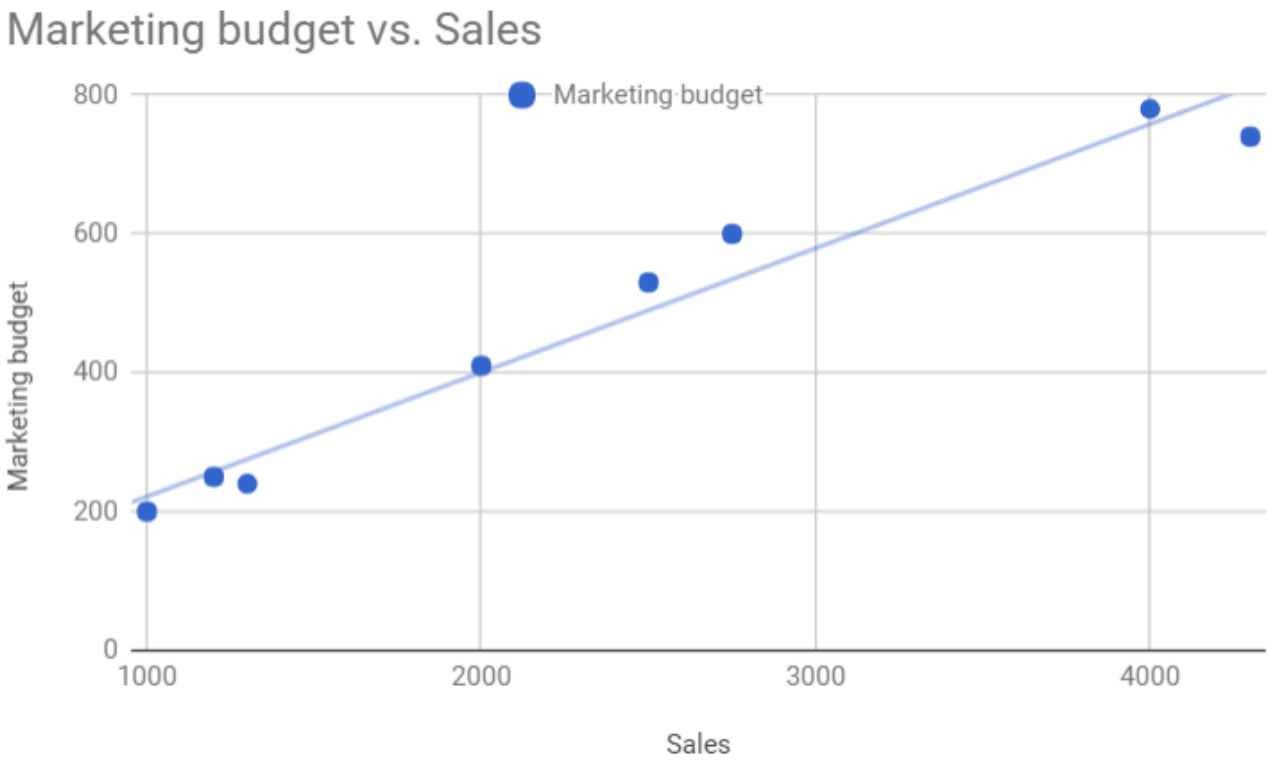

Figure 1. A linear regression

"Regular statistics" enables, among other things:

1. to find the numerical relation between the 2 series, based on a pre-established formal model (eg, ordinary least squares).

→ we see that sales are correlated with marketing spendings. It is likely that more marketing spending causes more sales.

2. to predict, based on this model:

→ by tracing the line further (using the formal model), we can predict the effect of more marketing spending

"Regular statistics" is advanced by scientists who:

1. are highly skilled in mathematics

→ their goal is to find the exact mathematical expression defining the situation at hand, under rigorous conditions

→ a key approach is **inference**: by defining a **sample of the data** of just the correct size, we can reach conclusions which are valid for the entire dataset.

2. have no training in computer science / software engineering

→ they neglect how hard it can be to to run their models on computers, in terms of calculations to perform. → since they focus on **sampling** the data, they are not concerned with handling entire datasets with related IT issues.

Machine learning does similar things to statistics, but in a slightly different way:

- there is an emphasis on getting the prediction right, not caring for identifying the underlying mathematical model
- the prediction needs to be achievable in the time available, with the computing resources available
- the data of interest is in a format / in a volume which is not commonly handled by regular statistics package (eg: images, Terabytes)

Machine learning is advanced by scientists who are typically:

- highly skilled in mathematics
- able to work with computer scientists, to take their constraints into account
- working in environments (industry, military, ...) where the operational aspects of the problem are key determinants (unstructured data, limits on computing resources)

(footnote: so machine learning, in my opinion, shares the spirit of "getting things done" as was operations research in the early days)

How does machine learning works, if it does not rely on modelling / sampling the data?

With 3 families of tricks:

b. The unsupervised learning approach

This designates all the methods which take a fresh dataset and find interesting patterns in it, **without training on previous, similar datasets**.

The analogy is with a person doing a task for the first time:

→ she learns a new thing by applying clever heuristics, without having been training on the task before.

Example: in your wedding, how to sit people with similar interests at the same tables?

The set up:

- a list of 100 guests, and 3 tastes you know they have for each of them
- 10 tables with 10 sits each.
- a measure of similarity between 2 guests: 2 guests have similarity of 0% if they share 0 tastes, 33% if they share 1 taste, 66% with 2 tastes in common, 100% with three matching interests.
- a measure of similarity at the level of a table: the sum of similarities between all pairs of guests at the table (45 pairs possible for a table of 10).

A possible solution using an unsupervised approach:

- on a computer, assign randomly the 100 guests to the 10 tables.
- for each table:
  - measure the degree of similarity of tastes for the table
  - exchange the sit of 1 person at this table, with the sit of a person at a different table.
  - measure again the degree of similarity for the table: if it improves, keep the new sits, if not, revert to before the exchange

And repeat for all tables, many times, until no exchange of sits improves the similarity. When this stage is achieved, we say the model has "**converged**".

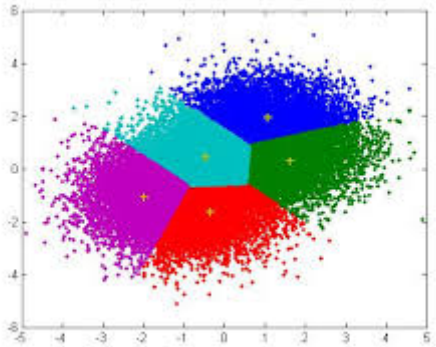

Figure 2. K-means, an unsupervised learning approach

c. The supervised learning approach

Take 50,000 or more observations, or data points, like:

**an image of a cat, with the caption "cat"

**an image of a dog, with the caption "dog"

**another image of a cat, with the caption "cat"

etc....

- you need 50,000 observations of this kind, or more! It is called the **training set**
- this is also called a **labelled dataset**, meaning that we have a label describing each of the observation.

The task is: if we give our computer a new image of a cat without a label, will it be able to guess the label "cat"?

The method:

- take a list of random coefficients (in practice, the list is a vector, or a matrix)
- for each of the 50,000 pictures of dogs and cats:
  - apply the coefficients to the picture at hand (let's say we have a dog here)
  - If the result is "dog", do nothing, it works!
  - If the result is "cat", change slightly the coefficients.
  - move to the next picture
- After looping through 50,000 pictures the parameters have hopefully adjusted and fine tuned. This was the **training of the model**.

Now, when you get new pictures (the **fresh set**), applying the trained model should output a correct prediction ("cat" or "dog").

Supervised learning is currently the most popular family of machine learning.

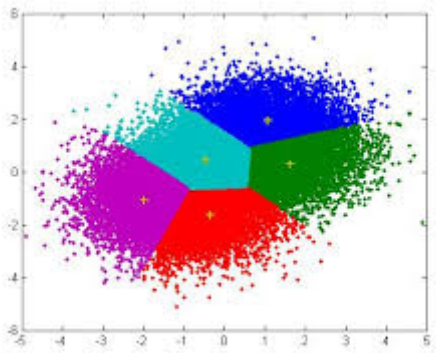

Figure 3. A hard test case for supervised learning

▶ https://www.youtube.com/watch?v=4HCE1P-m1l8 *(YouTube video)*

It is called **supervised** learning because the learning is very much constrained / supervised by the intensive training performed:

→ there is limited or no "unsupervised discovery" of novelty.

Important take away on the supervised approach: **collecting *large* datasets for training is key**

Without these data, no supervised learning.

d. The reinforcement learning approach

Take the case of a video game like Super Mario Bros. Goal of the task: collecting gold coins and completing the game.

- Starting point: Mario Bros is standing at the beginning of the game, doing nothing.
- Randomness is introduced: try something ("move right")
- The game ends (Mario moved right, gets hit by a ennemy)
- This negative result is stored somewhere (walking close to an ennemy = not good)
- Game starts over
- Randomness is introduced: try something different ("move right and make random jumps")
- The game ends (Mario moved right, jumped above the ennemy, collected gold coins with the jump, then got hit by a nasty cloud or fire when jumping)
- This new negative result is stored somewhere (walking close to an ennemy = not good, jumps near fire = not good, jumps = more gold coins)
- Game starts over
- Etc…

Note: reinforcement is both positive ("jumps help collect gold coins") and negative ("walking straight to an ennemy ends the game")

▶ https://www.youtube.com/watch?v=qv6UVOQ0F44 *(YouTube video)*

Reinforcement learning is perceived as corresponding to an important side of human learning / human intelligence (goal oriented, "trial and error").

These 3 families are called together **machine learning** because:

- the follow a looping process which is, really, an algorithm made to run on a computer, not a mathematical formula

→ a series of instructions ("do this, then that…") repeated quickly over and over again, which computers are good at performing.

- the solution is not found by applying a mathematical model, but emerges through a **learning**

process:

→ the coefficients get better and better at each loop of the algorithm.

2. Machine Learning and Data Science

Machine learning is a step in the longer chain of steps of data science.

"Doing data science" is very much discovering knowledge in data.

The process was formalized as kdd: "Knowledge Discovery in Databases":

[kdd] | *kdd.jpg*

Figure 4. KDD - knowledge discovery in databases

Machine learning is one of the techniques (along with traditional statistics) that intervenes at the step of "Data mining".

What makes data scientists important is that the steps of this kdd are highly interdependent.

You need indviduals or teams who are not just versed in data mining:

→ because the shape of the data at the collection stage has a huge influence on the kind of techniques, and the kind of software, that can be used to discover knowledge.

The skills of a data scientist are often represented as the meeting of three separate domains:

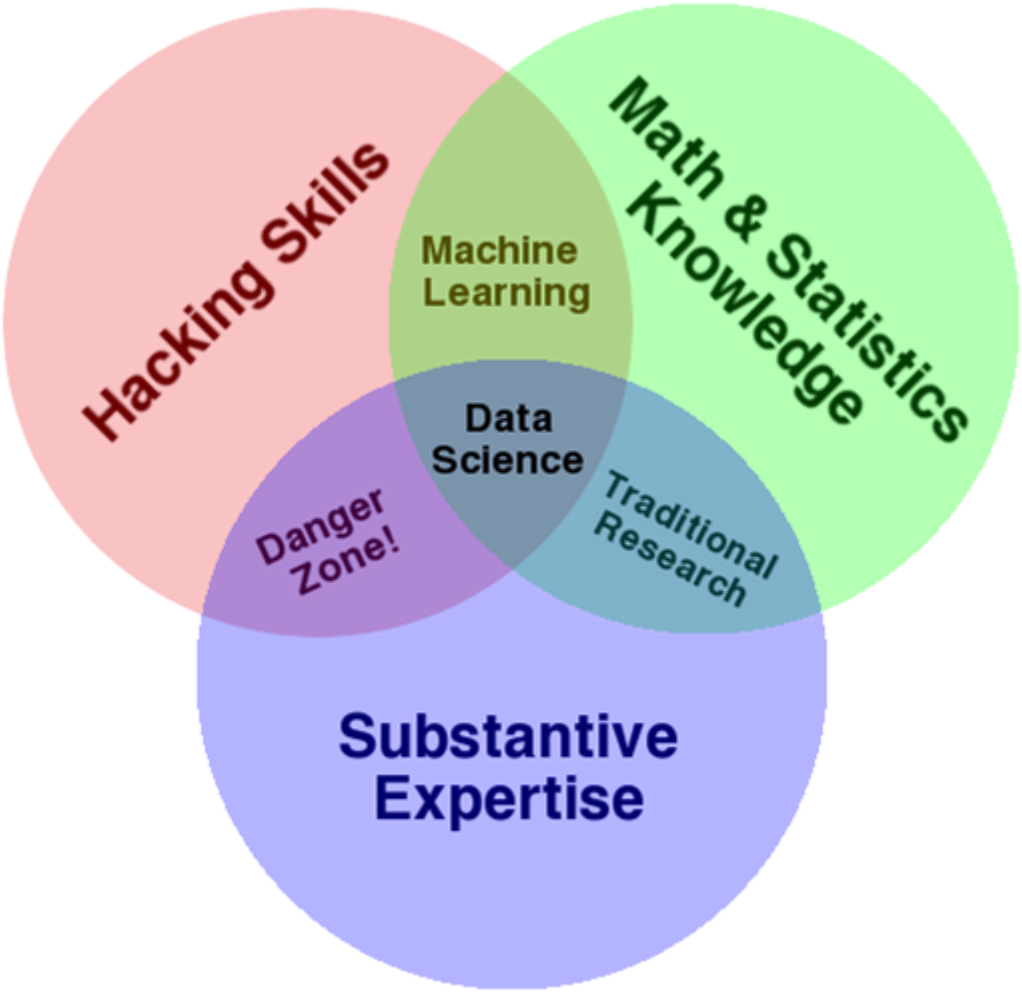

Figure 5. The Venn diagram of what is a data scientist

source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

3. Artificial intelligence

a. Weak vs Strong AI

We are currently at the stage of "weak AI", not being sure if "strong AI" will emerge.

Weak AI designates computer programs able to perform better than humans at complex tasks with a narrow focus (playing chess)

Weak AI is typically the results of applying machine learning techniques seen above.

Strong AI is an intelligence that would be general in scope, able to set its own goal, and conscious of itself. Nothing is close to that yet.

So AI is a synonymous with machine learning at the moment.

a. Two videos to understand AI further

Laurent Alexandre on the social and economic stakes of AI (in French):

▶ https://www.youtube.com/watch?v=rJowm24piM4 *(YouTube video)*

John Launchbury, the Director of DARPA's Information Innovation Office (I2O) in 2017:

▶ https://www.youtube.com/watch?v=-O01G3tSYpU *(YouTube video)*

The end

Find references for this lesson, and other lessons, here.

This course is made by Clement Levallois.

Discover my other courses in data / tech for business: http://www.clementlevallois.net

Or get in touch via Twitter: @seinecle