case_20.1

April 16, 2020

1 Data Science, Machine Learning, and Artificial Intelligence at a Glance

1.1 Introduction (5 mts)

What is data science? Machine learning? Artificial intelligence? These terms are all used interchangeably in our society, as if they are the same thing – but in reality they are quite different. Before we can answer those questions though, we have to tackle a much more fundamental one: What is data? Surprisingly, it is quite hard to come up with a simple definition for it.

1.1.1 Question: (5 mts)

What is "data"?

You can go onto Wikipedia or read books to get an answer to this question, but most of those sources will give you a very pedantic, unintuitive definition. Instead, we're going to go with the colloquial definition of data as "something whose value you care about". You won't find that in any formal treatment of the subject, but for now, it is good enough. Your name, age, and telephone number are data about you. Your bank savings, your address, and your parents' names are data that relate to you. We have data about everything, everywhere.

1.2 What is data science? (10 mts)

Now that we know what data is, we can now ask: "What is data science?" Science, in the language of the scientific method, is:

1. Formulating hypotheses, or guesses about how the world works, based on observations of the world around us
2. Validating or invalidating those hypotheses by conducting experiments

Unlike the pure sciences though, working with data doesn't necessarily require conducting experiments (although it could!). Rather, many times the data has already been collected and organized by someone else. So the scientific method, as applied to data, can be summarized as: **"Formulating hypotheses based on the world around us, then analyzing relevant data to validate or invalidate our hypotheses."**

1.2.1 Exercise 1: (5 mts)

Based on the above, which of the following can be described as reflective of the entire data science process? Select all that apply.

(a) Anecdotally noticing that Millennials seem to respond more positively to discussions of your firm's new product version that is in beta versus your existing one, setting up an A/B test funnelling Millenials equally to both versions, then conducting statistical significance tests on this data to verify that they prefer the new version.

(b) Observing that Uber pricing seems to be correlated to a small set of factors, obtaining open-source data on Uber pricing rates, then building a pricing model based on those factors and verifying that they explain most of the variation in rates.

(c) Converting images of crop circles in to structured pixels and storing them into a database for later use.

(d) Building an algorithm that allows a computer to recognize images of cats and dogs.

Answer. (a) and (b). Notice that all of these scenarios follow the "hypothesis - experiment/investigation - analysis" sequence described earlier. (c) is a **data engineering** problem; while not reflective of the entire data science process, this sort of manipulation of data to get it into a form suitable for analysis is a crucial part of the data science process. (d) is also not reflective of the entire process, as we are not doing any hypothesis testing, just building a model. We will revisit this soon enough.

Notice that data science is **NOT** what is often brought up in the media:

1. It is **NOT** computers recognizing images of cats and dogs
2. It is **NOT** IBM Watson screening human tissues for disease
3. It is **NOT** AlphaGo beating the world's top Go player

In fact, the VAST majority of uses of data are **NOT** like the above at all, and are instead much more similar to the traditional sciences. They are what we described in choices (a), (b), and (d) in Exercise 1 along with countless other similar types of scenarios.

1.3 What is machine learning? (10 mts)

However, choice (d) of Exercise 1, as well as examples 1 and 2 from the above list, do fall into the bucket of **machine learning**. What does this mean?

"Learn" means to "gain or acquire knowledge or skill in something via experience." So one could frame "machine learning" as "how a machine gains or acquires knowledge via experience." How does a machine gain experience? All machine inputs are essentially binary strings of 0s and 1s, which is really just – you guessed it – data! So machine learning is really just **how a computer acquires knowledge via data**.

Of course, this gives no insight into the "how" at all; it just says that there is soemething that is done with input data to generate this knowledge as an output. To make a math analogy, machine learning is some function f such that

$ knowledge = f(data)$

and other than that there are no other real stipulations on f! So f could be as mechanical as a simple mathematical function (say, the sum of all the data points) and qualify as machine learning. And in practice, this is what most of the common machine learning algorithms are, including:

1. Logistic regression
2. Random forests
3. Support vector machines
4. k - means clustering
5. Neural networks

(You will learn about all of these later in the course.) This may seem disappointing, given how the media hypes up "artificial intelligence" and makes it seem like there is something "smart" going on with machine learning, but in fact many mechanical methods satisfy the conditions required to be classified as machine learning. This doesn't mean these mechanical methods are limited in usefulness – in fact, they are quite powerful if used properly – but it does mean that they don't resemble anything that we would naturally associate with human-like intelligence.

1.3.1 Exercise 2: (5 mts)

Based on the above definition, which of the following tasks would likely involve the use of "machine learning"? Select all that apply.

(a) Building the model backing Apple iPhones' facial recognition system.

(b) Constructing the model backing Netflix's movie recommendation system based on your previous viewing activity.

(c) Investigating factors which affect Airbnb pricing and developing a pricing tool based on this analysis.

(d) Setting up an automated system to approve or reject mortgage loan applications.

Answer. All of the above.

(a) is similar to the task of recognizing cats and dogs in images, which as we have discussed is a machine learning problem.

(b) involves building a computer model that can learn your movie preferences based on your previous viewing activity; again, this fits into our previous definition of machine learning.

While (c) does not explicitly reference building a machine learning model, a pricing tool which takes into account all the factors affecting Airbnb pricing would likely be complex enough to benefit from the use of machine learning.

(d) also does not explicitly reference building a model, but there are many parts to a mortgage loan application, and therefore likely many factors that are relevant in determining approval. Thus, any automated system would likely require a complex model that takes all of these factors into account. Such a model would benefit from incorporating machine learning.

Thus, machine learning can be part of conducting data science. That is, **data science is fundamentally a process, while machine learning is a tool that can be immensely useful in conducting the data science process**.

1.4 What is artificial intelligence? (30 mts)

But the elephant is still in the room: even though some mechanical, "dumb" methods may qualify as machine learning, this doesn't exclude human-like, "smart" methods from being classified as such either. And semantically, this is completely true – it doesn't, yet people have chosen to name it something else entirely: **artificial intelligence**.

But why? Why give "smart" methods an entirely different name if they can also fall under the bucket of machine learning? That is the question we will explore for the remainder of this module.

Let's start by taking a look at an iconic demonstration of this so-called intelligence: AlphaGo beating the world's top human Go player.

[8]:
```
from IPython.display import HTML
```

[9]:
```
HTML('<iframe width="560" height="315" src="https://www.youtube.com/embed/
↪8tq1C8spV_g?rel=0&amp;controls=0&amp;showinfo=0" frameborder="0"␣
↪allowfullscreen></iframe>')
```

[9]:
```
<IPython.core.display.HTML object>
```

Quite impressive! But does this feat alone prove that a machine exhibits human-like intelligence?

1.4.1 Question: (5 mts)

What, to you, counts as "artificial intelligence"? What demonstrations of aptitude would, beyond a shadow of a doubt, convince you that something is as intelligent as us humans are?

You may have found that it was quite hard to come up with answers to the second part of the previous question, and that most ideas you had either: 1) seemed like they could well have a "dumb" mechanical solution as in the previous discussion of machine learning, despite seeming impressive at first; or 2) actually made you question how unique human intelligence really is and whether virtually all of what we do could be reduced to such "dumb" mechanical methods. We'll explore both ideas below.

1.4.2 What can machines do and not do? What about us? (5 mts)

So, is there any sensible test we could use to determine if something is as intelligent as a human? There have been many proposals over time. The most famous aptitude test developed was the **Turing test**, named after the English mathematician and famous World War II cryptographer Alan Turing. In this exam, there is a human evaluator and two conversation partners: one machine and one human. The evaluator would conduct a conversation with each through a text-only channel. If the evaluator cannot reliably tell the machine from the human, the machine is said to have passed the test.

Turing did not explicitly state that his test could be used as a measure of intelligence, but nonetheless many who came after him thrust his test into the limelight. Of course, the corollary of this is that if a computer can converse like a human, then it is effectively as intelligent as a human.

1.4.3 Question: (5 mts)

What are some shortcomings of this proposal? What insight does this lend into the criteria that a rigorous test of human-like intelligence would have to satisfy?

In addition to the flaws with the Turing test (and in fact with almost any other test you can likely come up with), this brings to light one of our society's unhealthy obsessions when it comes to the field of artificial intelligence – its singular focus on mimicking human intelligence via machines. But what if machine intelligence is fundamentally different (note: not worse) than human intelligence? What if machines are more "intelligent" about certain things than we are, and vice versa?

1.4.4 Question: (10 mts)

Brainstorm with a partner, and then we will discuss aloud in class:

1. What are some things that machines can already do better than we can? What specifically about them allows them to do these better?
2. What are some things that we can do better than machines? Do you think this advantage will likely be sustained over time? Why?

1.5 Conclusions & Takeaways (5 mts)

In this discussion, you learned what "data science" and "machine learning" really are, in contrast to the misleading connotations that they are often given in public discussion. You also learned that "artificial intelligence" is a very murky term – nobody really knows what it is, and it's unclear if in its current focus on imitating human intelligence, that it is even a good use of our time and effort.

Throughout this program, we will focus primarily on data science and machine learning, not so much artificial intelligence. Yet the philosophical questions surrounding artificial intelligence are fascinating, and we encourage you to continue pondering them as you become more involved in this new and exciting field.

[ ]: