Introduction to Machine Learning

APT 3025

SPRING 2022

Overview

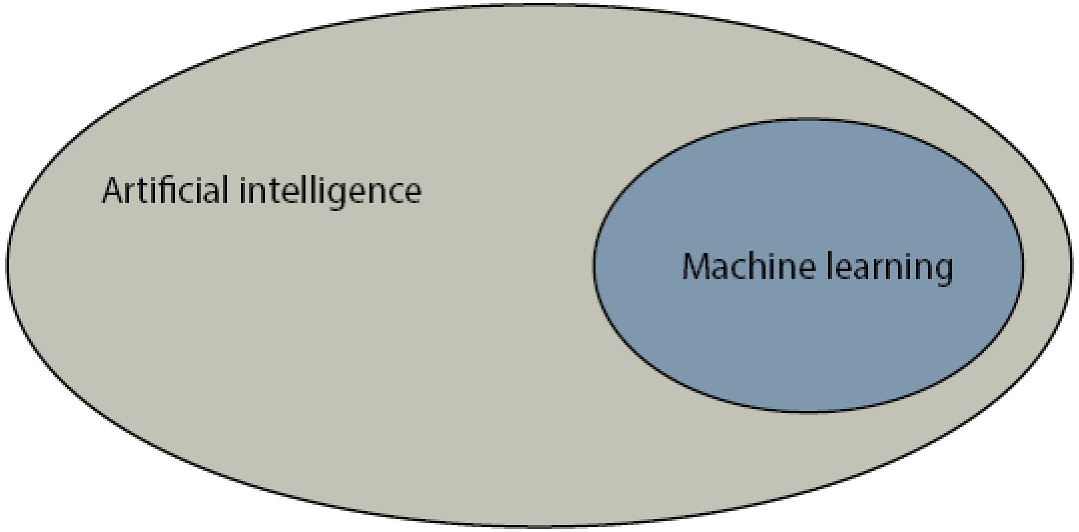
- What is machine learning?
- What will this course cover?
- What is artificial intelligence and how does it differ from machine learning?
- What is deep learning
- How do humans think, and how can we inject those ideas into a machine?
- Some basic machine learning examples in real life

Artificial Intelligence

- To define machine learning, first let’s define a more general term: artificial intelligence.
- Artificial intelligence is the set of all tasks in which a computer can make decisions.
- These decisions allow a computer to solve a problem by itself, such as: driving a car, finding a route between two points, deciding whether an incoming email is spam, predicting the price of a house, or recommending a movie.

What is Machine Learning?

- We can define machine learning as the set of all tasks in which a computer can make decisions *based on data*.



How humans make decisions

- Humans make decisions in the following two ways:
  - By using logic and reasoning
  - By using our experience
- Imagine that we are trying to decide what car to buy.
- We can look carefully at the features of the car, such as price, fuel consumption, and navigation, and try to figure out the best combination of them that fits our budget.
- This is using logic and reasoning.

How humans make decisions

- If instead you ask all your friends what cars they own, and what they like and dislike about them
- Or perhaps you have driven two types of cars in the past, type A and type B. You liked type A and hated type B. Now you want a car as that is similar to type A and different from type B.
- In these cases, you collect information and use that to decide, that is, you are using experience (in the first case, your friends’ experiences, an in the second, your own).
- Machine learning refers to this method: making decisions using experience. In computer lingo, the term for experience is data.

Instructions vs. examples

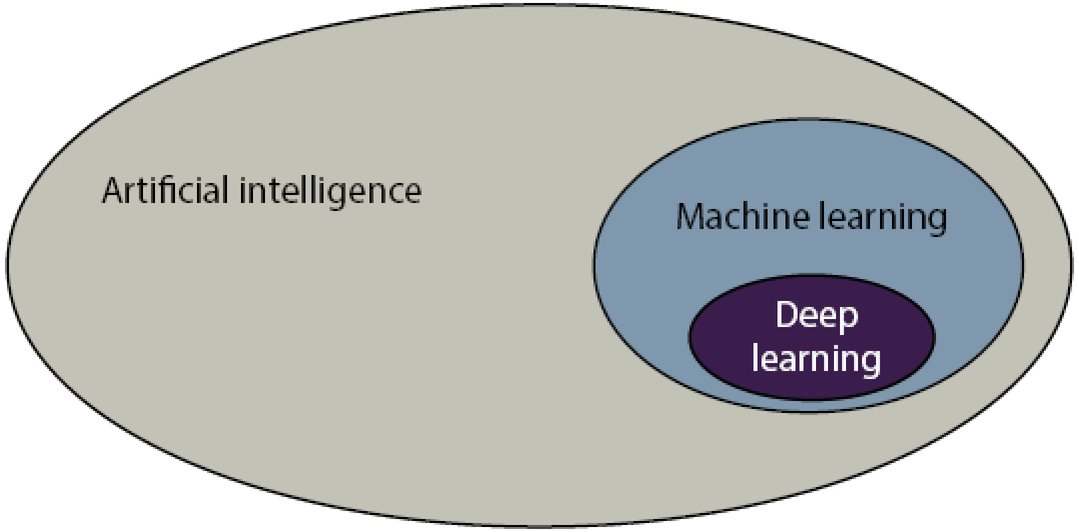
- Normally, to get a computer to perform a task, we have to write a program, namely, a set of instructions for the computer to follow.
- This process is good for simple tasks, but some tasks are too complicated for this framework.
- For example, consider the task of identifying if an image contains an apple. If we start writing a computer program to develop this task, we quickly find out that it is hard.
- A child does not learn what an apple is by using a set of instructions. S/he learns by seeing many examples again and again, and hearing adults label these examples “apple”.

Learning from examples

- In machine learning, we show the computer many images, and we tell it which ones contain an apple (that constitutes our data).
- We repeat this process until the computer learns the right patterns and attributes that constitute an apple.
- When we feed the computer a new image, it can use these patterns to determine whether the image contains an apple.
- Of course, we still need to program the computer so that it learns these patterns. For that, we have several techniques, which we will discuss in this course.

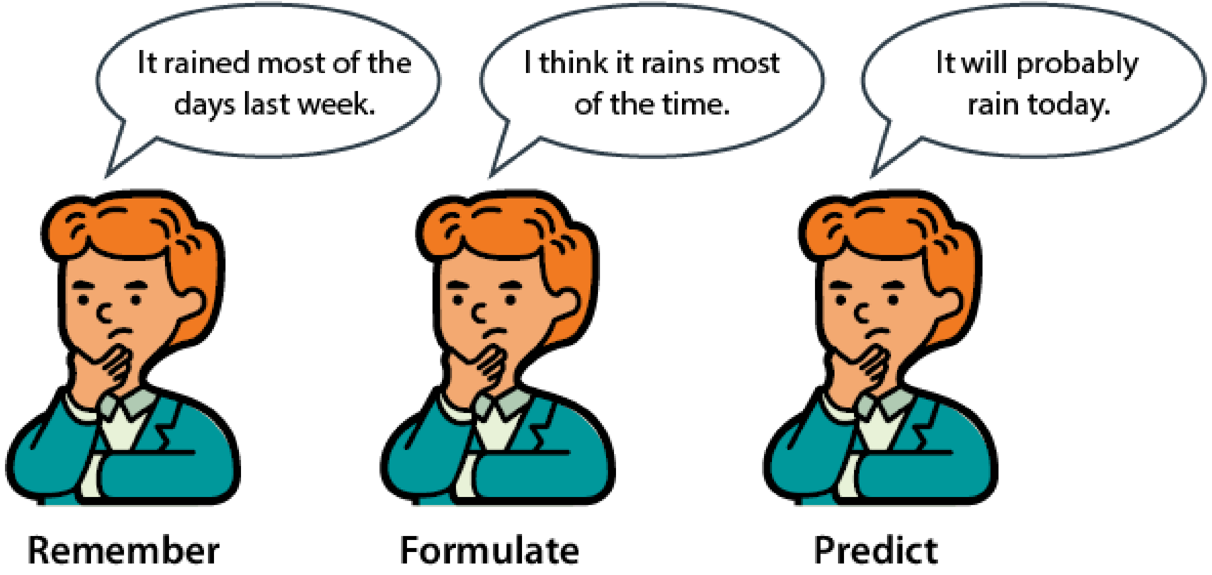
Deep Learning

- Deep learning is a type of machine learning that uses a particular technique known as neural networks.
- Deep learning has gained a lot of attention because of its success in solving problems that were impossible for a long time.



The remember-formulate-predict framework

- When we, as humans, need to make a decision based on our experience, we normally use the following framework:
  - We remember past situations that were similar.
  - We formulate a general rule.
  - We use this rule to predict what may happen in the future.


It rained most of the days last week.
I think it rains most of the time.
It will probably rain today.
Remember
Formulate
Predict

Some machine learning terminology

- In machine learning, the way the computer solves a problem is by using the data to build a *model*.
- A *model* is a set of rules that represent our data and can be used to make predictions.
- An *algorithm* is the process that we used to build the model.
- In the example, the process is simple: we looked at how many days it rained and realized it was the majority. That is our model.
- An *algorithm* is a procedure, or a set of steps, used to solve a problem or perform a computation.
- The goal of a *machine learning algorithm* is to build a *model*.

Spam Detection

- *Spam* is the common term used for junk or unwanted email, such as chain letters, promotions, and so on.
- The term *ham* is used to refer to non-spam emails.
- In this example, our friend Bob likes to send us email. A lot of his emails are spam, in the form of chain letters. We are starting to get a bit annoyed with him. It is Saturday, and we just got a notification of an email from Bob. Can we guess if this email is spam or ham without looking at it?

A very simple model

- Let’s use the remember-formulate-predict method.
- **Remember**: of the last 10 emails we got from Bob, 6 were spam and 4 were ham. That is our data.
- From this information, we can **formulate** the following model:
- **Model 1**: Six out of every 10 emails that Bob sends us are spam.
- This model may not be that good, but it’s the best we can get from thee data we have.
- **Predict**: Given this model, it is more likely that this incoming email is spam rather than ham.

Improvement with more data

- Additional data: Let’s see **when** the emails were sent.

Email No.	Day	Type of email
1	Monday	Ham
2	Tuesday	Ham
3	Saturday	Spam
4	Sunday	Spam
5	Sunday	Spam
6	Wednesday	Ham
7	Friday	Ham
8	Saturday	Spam
9	Tuesday	Ham
10	Thursday	Ham

A better model

- We notice a pattern. It seems that every email Bob sent during the week is ham, and every email he sent during the weekend is spam.
- So, we can formulate a more educated rule, or model, as follows:
- **Model 2**: Every email that Bob sends during the week is ham, and those he sends during the weekend are spam.
- Now suppose today is Saturday. Then we can predict with great confidence that the email Bob just sent is spam. We make this prediction, and without looking, we send the email to the trash.

Model 3

- Suppose it turns out we missed Bob’s birthday party because we trashed the invitation! Our model made a mistake.
- Let’s dig deeper into the data and see if there’s something else we need to remember.

Email Size (KB)	Type of email
1	Ham
2	Ham
16	Spam
20	Spam
18	Spam
3	Ham
5	Ham
25	Spam
1	Ham
3	Ham

Model 3

- It seems that the large emails tend to be spam, whereas the smaller ones tend to be ham.
- So, we can formulate the following rule:
- **Model 3**: Any email of size 10 KB or larger is spam, and any email of size less than 10 KB is ham.
- Now that we have formulated our rule, we can make a **prediction**. We look at the email we’ve just received from Bob, and the size is 19 KB. So, we conclude that it is spam.

Features

- To make our predictions, we used the day of the week and the size of the email. These are examples of features.
- A feature is one of the most important concepts in machine learning.
- A **feature** is any property or characteristic of the data that the model can use to make predictions.
- Features describe the things we want to make predictions about. For example, the size and date describe the email.
- What are some other features that can describe an email?

Model 4 and 5

- Our two classifiers were good, because they rule out large emails and emails sent on the weekends. Each one of them uses exactly one of these two features. But what if we wanted a rule that worked with both features? Rules like the following may work:
- **Model 4**: If an email is larger than 10 KB or it is sent on the weekend, then it is classified as spam. Otherwise, it is classified as ham.
- **Model 5**: If the email is sent during the week, then it must be larger than 15 KB to be classified as spam. If it is sent during the weekend, then it must be larger than 5 KB to be classified as spam. Otherwise, it is classified as ham.

An even more complex model

- **Model 6**: Consider the number of the day, where Monday is 0, Tuesday is 1, Wednesday is 2, Thursday is 3, Friday is 4, Saturday is 5, and Sunday is 6. If we add the number of the day and the size of the email (in KB), and the result is 12 or more, then the email is classified as spam. Otherwise, it is classified as ham.
- We can continue formulating more models with additional layers of complexity.
- We will need a way to choose the best model, a way to evaluate models so as to compare them.

How a computer develops a model

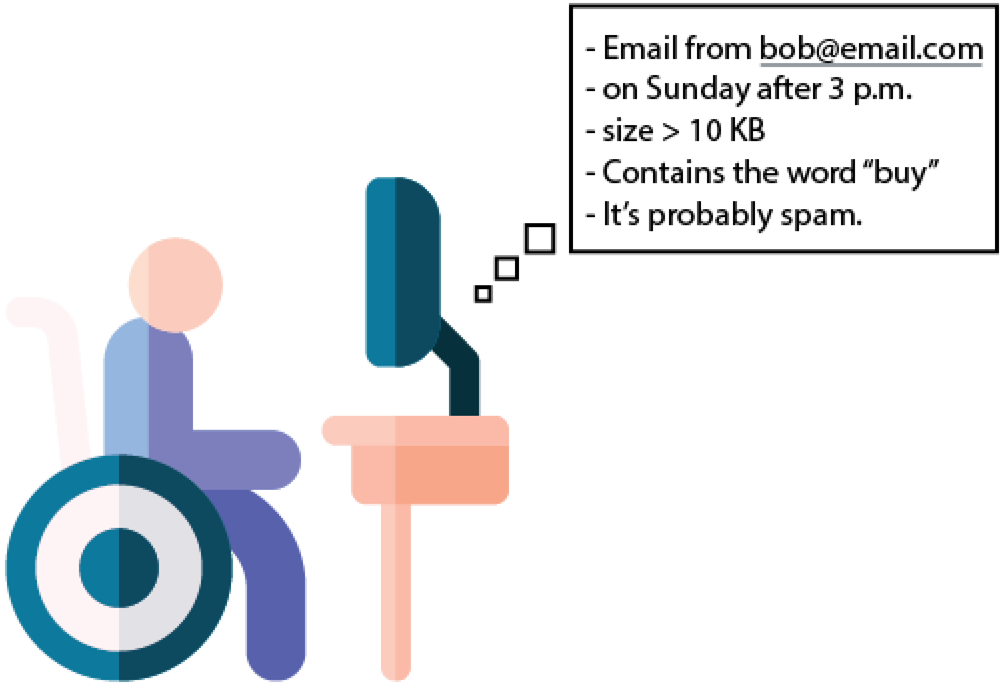
- **Remember**: Look at a huge table of data.
- **Formulate**: Create models by going through many rules and formulas, and check which model fits the data best.
- **Predict**: Use the model to make predictions about future data.
- These are the same steps we followed, but the computer can build models quickly by going through many formulas and combinations of rules until it finds one that fits the existing data well.
- In addition, a computer can process massive amounts of data in order to generate a model.

Model 7

- A computer could build a spam classifier with features such as the sender, the date and time of day, the number of words, the number of spelling mistakes, and the appearances of certain words such as *buy* or *win*.
- **Model 7**
  - If the email has two or more spelling mistakes, then it is classified as spam.
  - If it has an attachment larger than 10 KB, it is classified as spam.
  - If the sender is not in our contact list, it is classified as spam.
  - If it has the words buy and win, it is classified as spam.
  - Otherwise, it is classified as ham.

Model 8

- We can have an even more complex model.
- **Model 8**: If (size) + 10 (number of spelling mistakes) – (number of appearances of the word “mom”) + 4 (number of appearances of the word “buy”) > 10, then we classify the message as spam. Otherwise, we classify it as ham.



- Email from bob@email.com
- on Sunday after 3 p.m.
- size > 10 KB
- Contains the word “buy”
- It’s probably spam.

Generalisation

- Which is the best model?
- It may seem that the one that fits the data best should be considered the best model.
- Actually, we want one that makes the most accurate predictions for new data (i.e. data that was not used to formulate the model).
- This is known as generalisation. We want a model that generalizes well.

Practical Work

- We will use the following software tools and libraries:
    1. Python
    2. Scikit-learn
    3. Pandas
    4. Matplotlib
    5. TensorFlow
    6. Keras
    7. Jupyter
- Google Colaboratory (Colab) is an online platform equipped with all the tools and libraries we will need.
- You may also install Anaconda on your computer. It comes with everything we'll need.

Machine Learning Applications

- Machine learning is not new.
- Successful applications that have existed for decades are:
  - Optical character recognition
  - Handwriting recognition
  - License plate recognition
  - Spam filtering
- Machine learning has achieved astounding advances within the last decade, solving previously intractable problems, e.g.:
  - Speech recognition
  - Image search