Supervised Machine Learning: Regression and Classification

What is machine learning?

“Field of study that gives computers the ability to learn without begin explicitly programmed” - Arthur Samuel (1959)

- The amazing thing about this program was that Arthur Samuel himself wasn't a very good checkers player.
- What he did was he had programmed the computer to play maybe tens of thousands of games against itself.
- By watching what social support positions tend to lead to wins and what positions tend to lead to losses the checkers plane program learned over time what are good or bad suport positions by trying to get a good and avoid bad positions, this program learned to get better and better at playing checkers because the computer had the patience to play tens of thousands of games against itself.
- It was able to get so much checkers playing experience that eventually it became a better checkers player than also, Samuel himself.

Quizz:

If the checkers program had been allowed to play only ten games (instead of tens of thousands) against itself, a much smaller number of games, how would this have affected its performance?

◯ Would have made it better

◉ Would have made it worse (choose this)

Arthur Samuel's definition was a rather informal one but in the next two videos, we'll dive deeper together into what are the major types of machine learning algorithms?

Machine learning algorithms

- Supervised learning (course 1, 2)

    - supervised learning is the type of machine learning that is used most in many real-world applications
    - and has seen the most rapid advancements and innovation.
- Unsupervised learning (course 3)
- Recommender systems
- Reinforcement learning

The other thing we're going to spend a lot of time on in this specialization is practical advice for applying learning algorithms.

- Teaching about learning algorithms is like giving someone a set of tools and equally important, so even more important to making sure you have great tools is making sure you know how to apply them

Supervised learning

- I think 99 percent of the economic value created by machine learning today is through one type of machine learning, which is called supervised learning.

Definition

- Supervised machine learning or more commonly, supervised learning, refers to algorithms that learn x to y or input to output mappings.
- The key characteristic of supervised learning is that you give your learning algorithm examples to learn from.
- That includes the right answers, whereby right answer, I mean, the correct label y for a given input x, and is by seeing correct pairs of input x and desired output label y that the learning algorithm eventually learns to take just the input alone without the output label and gives a reasonably accurate prediction or guess of the output.

Supervised learning

Learns from being given "right answers"

Example:

Input(x)	Output(Y)	Application
email →	spam ? (0 / 1)	spam filtering
audio →	text transcripts	Speech recognition
English →	Spanish	Machine translation
ad, user info →	clicks? (0/1)	online advertising
image, radar info →	position of other cars	Self-driving car
Image of phone	defect? (0/1)	visual inspection

After the model has learned from these input, output, or x and y pairs, they can then take a brand new input x, something it has never seen before, and try to produce the appropriate corresponding output y.

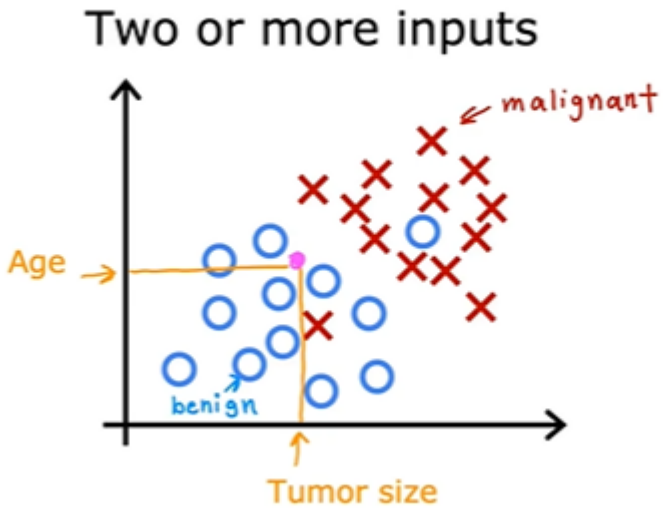
Let's dive more deeply into one specific example.

Regression algorithm (a type of supervised learning) :

I mean we're trying to predict a number from infinitely many possible numbers such as the house prices in our example, which could be 150,000 or 70,000 or 183,000 or any other number in between

Classification: Breast cancer detection

- In classification problems you can also have more than two possible output categories.
- And by the way in classification, the terms output classes and output categories are often used interchangeably. So what I say class or category when referring to the output, it means the same thing. So to summarize classification algorithms predict categories.
- Categories don't have to be numbers. It could be non numeric for example, it can predict whether a picture is that of a cat or a dog. And it can predict if a tumor is benign or malignant. Categories can also be numbers like 0, 1 or 0, 1, 2. But what makes classification different from regression when you're interpreting the numbers is that classification predicts a small finite limited set of possible output categories such as 0, 1 and 2 but not all possible numbers in between like 0.5 or 1.7.



What in this new dataset we're going to use circles to show patients whose tumors are benign and crosses to show the patients with a tumor that was malignant. So when a new patient comes in, the doctor can measure the patient's tumor size and also record the patient's age. And so given this, how can we predict if this patient's tumor is benign or malignant? Well, given the day said like this, what the learning algorithm might do is find some boundary that separates out the malignant tumors from the benign ones. So the learning algorithm has to decide how to fit a boundary line through this data. The boundary line found by the learning algorithm would help the doctor with the diagnosis. In this case the tumor is more likely to be benign.

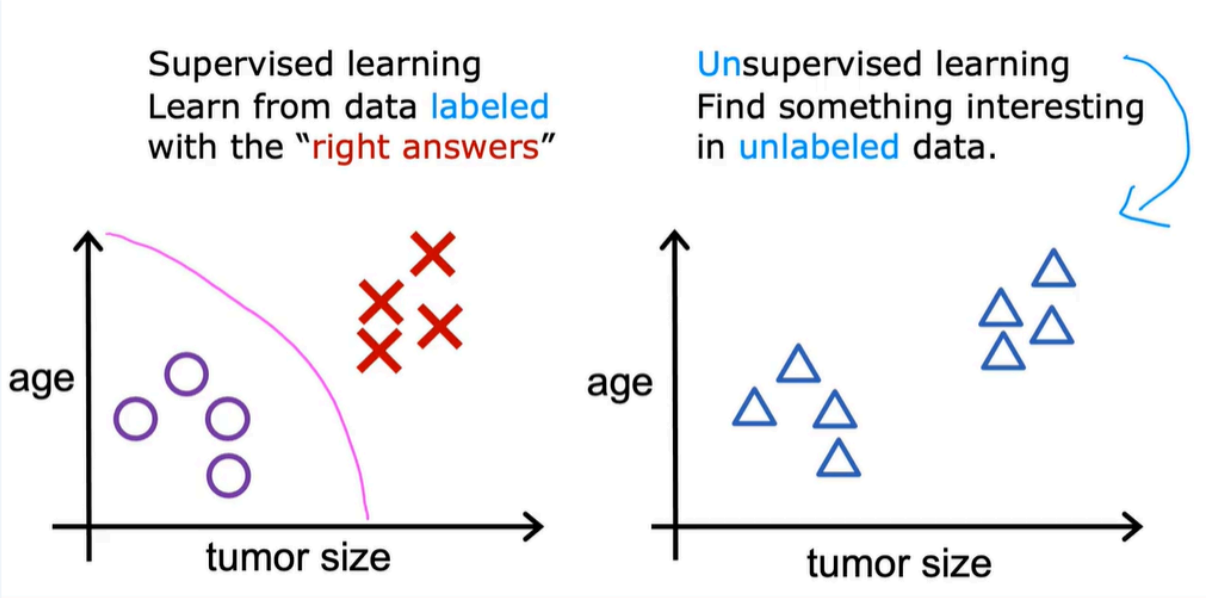
- My friends who worked on breast cancer detection use many additional inputs, like the thickness of the tumor clump, uniformity of the cell size, uniformity of the cell shape and so on

Summary

Learns from being given “right answers”

Regression	Classification
Predict a number	Predict categories
Infinitely many possible outputs	Small number of possible outputs

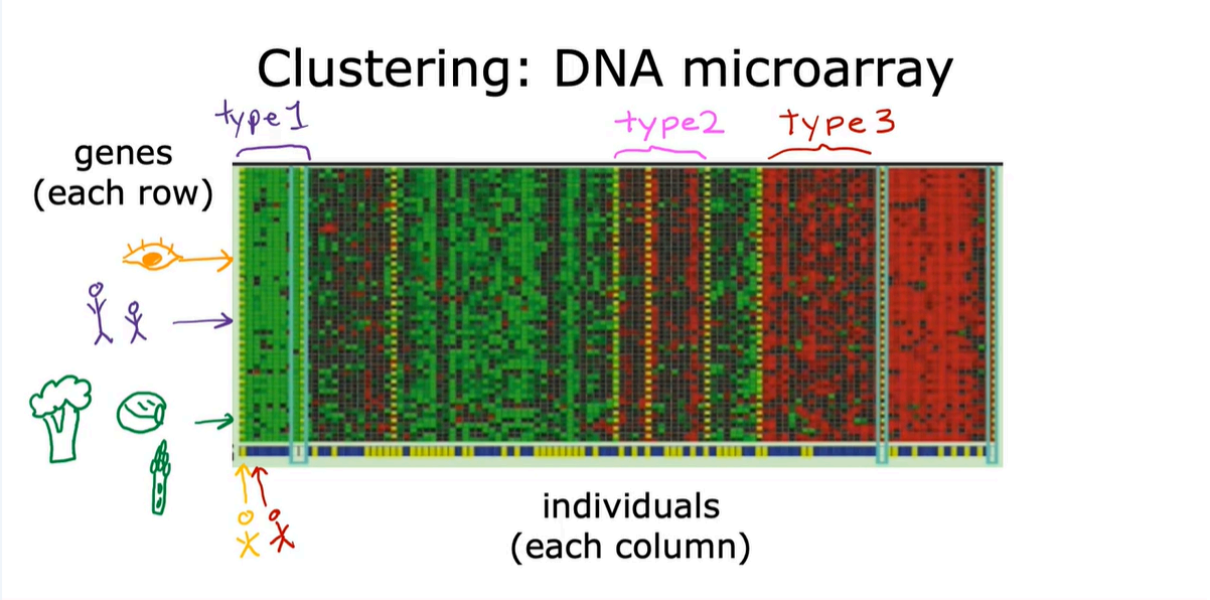
Unsupervised learning



- We're not asked to diagnose whether the tumor is benign or malignant, because we're not given any labels. Why in the dataset, instead, our job is to find some structure or some pattern or just find something interesting in the data. This is unsupervised learning, we call it unsupervised because we're not trying to supervise the algorithm.
- To give some quote right answer for every input, instead, we asked the our room to figure out all by yourself what's interesting.
- Or what patterns or structures that might be in this data, with this particular data set. An unsupervised learning algorithm, might decide that the data can be assigned to two different groups or two different clusters.
- An unsupervised learning algorithm, might decide that the data can be assigned to two different groups or two different clusters. And so it might decide, that there's one cluster what group over here, and there's another cluster or group over here. This is a particular type of unsupervised learning, called a clustering algorithm. Because it places the unlabeled data, into different clusters and this turns out to be used in many applications.

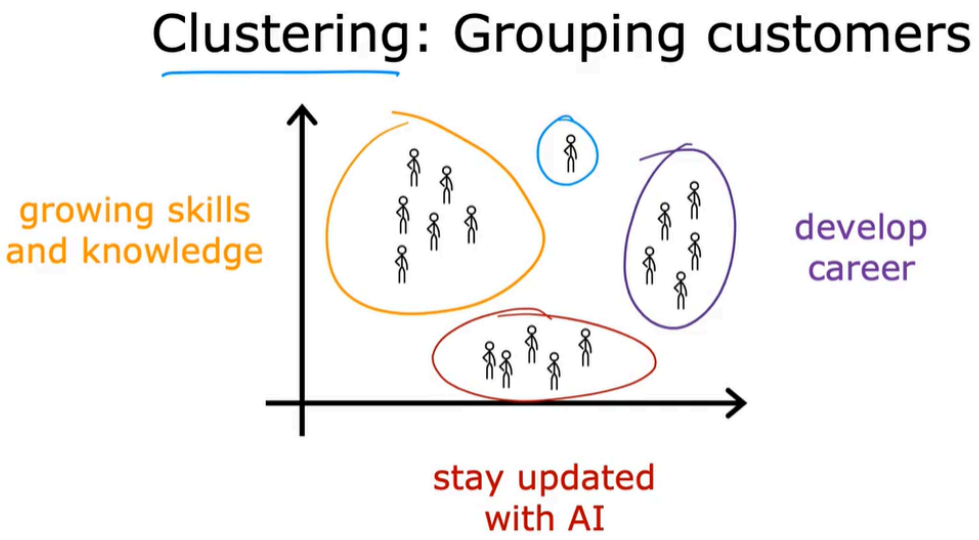
Notice that the word panda appears here here, here, here and here and notice that the word twin also appears in all five articles. And the word Zoo also appears in all of these articles, so the clustering algorithm is finding articles. All of all the hundreds of thousands of news articles on the internet that day, finding the articles that mention similar words and grouping them into clusters.

- And there are so many news stories, it just isn't feasible to people doing this every single day for all the topics that use covers. Instead the algorithm has to figure out on his own without supervision, what are the clusters of news articles today. So that's why this clustering algorithm, is a type of unsupervised learning algorithm.



This is unsupervised learning, because we're not telling the algorithm in advance, that there is a type one person with certain characteristics.

- Or a type two person with certain characteristics, instead what we're saying is here's a bunch of data. I don't know what the different types of people are but can you automatically find structure into data. And automatically figure out whether the major types of individuals, since we're not giving the algorithm the right answer for the examples in advance. This is unsupervised learning,



so to summarize a clustering algorithm. Which is a type of unsupervised learning algorithm, takes data without labels and tries to automatically group

them into clusters. And so maybe the next time you see or think of a panda, maybe you think of clustering as well. And besides clustering, there are other types of unsupervised learning as well.

Unsupervised learning

- Data only comes with inputs x, but not output labels y
- and the algorithm has to find some structure or some pattern or something interesting in the data

Clusering

- Group similar data points together

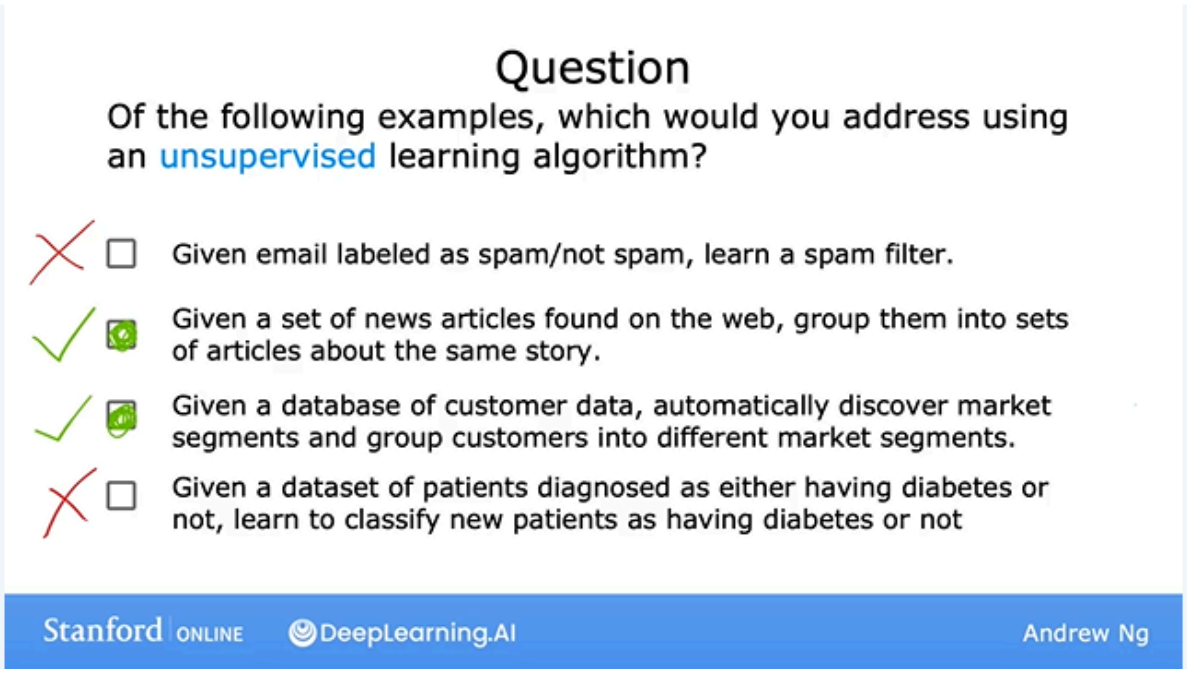
Anomaly detection

- Find unusual data points
- This turns out to be really important for fraud detection in the financial system, where unusual events, unusual transactions could be signs of fraud and for many other applications.

Demensionality reduction

- This lets you take a big data-set and almost magically compress it to a much smaller data-set while losing as little information as possible.
- Compress data using fewer numbers

In case anomaly detection and dimensionality reduction don't seem to make too much sense to you yet. Don't worry about it. We'll get to this later in the specialization.


Question
Of the following examples, which would you address using an unsupervised learning algorithm?
Given email labeled as spam/not spam, learn a spam filter.
Given a set of news articles found on the web, group them into sets of articles about the same story.
Given a database of customer data, automatically discover market segments and group customers into different market segments.
Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not
Stanford ONLINE
DeepLearning.AI
Andrew Ng