UNSUPERVISED LEARNING

1. Unsupervised learning

00:00 - 00:02

Great job!

2. Unsupervised learning

00:02 - 00:05

Let's switch over to unsupervised learning.

3. Unsupervised learning

00:05 - 00:33

Unsupervised learning is quite similar to supervised learning, except it doesn't have a target column - hence the unsupervised part. So what's the point then? Unsupervised learning learns from the dataset, and tries to find patterns. That's the reason this technique is so interesting and powerful: we can find insights without knowing much about our dataset.

4. Applications

00:33 - 00:42

Unsupervised learning has different applications. We'll focus on clustering, anomaly detection, association.

5. Clustering

00:42 - 00:57

Clustering consists in identifying groups in your dataset. The observations in these groups share stronger similarities with members of their group, than with members of other groups.

6. Clustering example

00:57 - 01:06

For example, say we have a dataset with six observations. What clusters would the algorithm detect?

7. Species cluster

01:06 - 01:13

Well, it depends. It may come up with two groups: dogs and cats.

8. Color cluster

01:13 - 01:20

Or, it might make four groups by color: black, grey, white and brown.

9. Origin cluster

01:20 - 01:49

Or, it may find origin groups: the top row originate from Europe, while the bottom row are from Japan. In these examples, I've told you what each group represents. However, you usually don't know what differentiates your clusters in real life. Your model won't tell you why or how it decided on these clusters. It's up to you to investigate and find out.

10. Clustering models

01:49 - 02:22

Some clustering models, like K-Means, require you to specify in advance the number of clusters you would like to identify. Others, like DBSCAN, or - get ready - "Density-based spatial clustering of applications with noise", don't require you to specify the number of clusters in advance. Instead, they require you to define what constitutes a cluster, like the minimum number of observations in one cluster.

11. Iris table

02:22 - 02:45

Let's say we have flowers of unknown species. All we have is their petal width and length. See the difference with a classification problem? Here, we don't have a column with labels of the species. We don't know which species we're dealing with or even how many there are.

12. K-Means with 4 clusters

02:45 - 03:00

If we hypothesize there are 4 species, we can use a K Means and require 4 different clusters. It will result in this clustering.

13. K-Means with 3 clusters

03:00 - 03:03

If we hypothesize there are 3 species, we require 3 different clusters.

14. Ground truth

03:03 - 03:13

These clusters are actually correct, as there are three different species in the dataset: Setosa, Virginica and Versicolor.

15. Anomaly detection

03:13 - 03:17

Let's now talk about anomaly detection.

16. Detecting outliers

03:17 - 03:28

Anomaly detection is all about detecting outliers. Outliers are observations that strongly differ from the others.

17. Outliers

03:28 - 03:50

On this picture, all of our points are grouped in the bottom left, except for one in the top right. This point is an outlier. It turns out that this point is the sum total of the other observations. The total row wasn't removed before plotting the data.

18. Removing outliers

03:50 - 04:13

In this case, the outlier can be removed. With two dimensions, it's easier to find outliers with our naked eye. Try finding outliers in 3 dimensions; that might be doable. How about 4, 10, 20,100? That's why we need unsupervised learning algorithms.

19. Some anomaly detection use cases

04:13 - 04:33

In our example, the outlier was an error. But detecting outliers can help find which devices fail faster or last longer, which fraudsters trick the protection systems in place, or which patients surprisingly resist a fatal disease.

20. Association

04:33 - 04:40

Let's end with association, which consists in finding relationships between observations.

21. Association

04:40 - 05:06

In other words, it's about finding events that happen together. It's often used for market basket analysis, which is just a fancy expression to state "Which objects are bought together?" For example, people who buy jam are likely to buy bread, people who buy beer are likely to buy peanuts, and people who buy wine are likely to buy cheese.

22. Let's practice!

05:06 - 05:11

All right, let's check your understanding.