Clustering

1. Clustering

00:00 - 00:10

Previously, we learned how to use Supervised Learning to make predictions based on labeled data. In this lesson, we'll cover another subset of machine learning called clustering.

2. What is clustering?

00:10 - 00:31

Clustering is a set of machine learning algorithms that divide data into categories, called clusters. Clustering can help us see patterns in messy datasets. Machine Learning Scientists use clustering to divide customers into segments, images into categories, or behaviors into typical and anomalous.

3. Supervised vs. unsupervised machine learning

00:31 - 00:58

Clustering is part of a broader category within Machine Learning called "Unsupervised Learning". Unsupervised Learning differs from Supervised Learning in the structure of the training data. While Supervised Learning uses data with features and labels, Unsupervised Learning uses data with only features. This makes Unsupervised Learning, and clustering, particularly appealing: you can use it even when you don't know much about your dataset.

4. Case study: discovering new species

00:58 - 01:18

Let's say you are a botanist and you've been doing field work on a previously unexplored island. Notably, you have several observations on these flowers you've never seen before. You believe you might have discovered a couple new flower species, but you're not sure how many and how to classify each flower. Let's see if clustering can help.

5. Defining features

01:18 - 01:40

The first step is finding features. Luckily, you've been meticulous in your data gathering and measured over 100 flowers. We can use your measurements as features for our model. This is indeed an unsupervised learning problem because we have features but we're not sure what species each flower belongs to or even how many new species there are!

6. Defining number of clusters

01:40 - 02:01

Some clustering algorithms need us to define how many clusters we want to create. The number of clusters we ask for greatly affects how the algorithm will segment our data. Here's

our flower data graphed over three features: petal width, sepal length, and number of petals on the x,y, and z axes, respectively.

7. Comparing number of clusters

02:01 - 02:17

Here is how the algorithm divides the data if we ask for two clusters. One color represents on cluster, in our case, one new flower species. And here is how it divides the same data if we ask for three clusters.

8. Comparing number of clusters

02:17 - 02:31

And these are the results when we ask for four and eight clusters. We can tell intuitively that eight is probably too many clusters, because there aren't as many clear cut areas in our graph.

9. Comparing number of clusters

02:31 - 03:02

Clustering won't tell us exactly how many clusters we have, but it can help us make an informed decision. In your case, it seems like you've found three or four new species. Having a strong hypothesis about our data helps us get better results from the clustering algorithm. For example, you may know from your experience as a botanist that petal width usually has wide variance within a species and shouldn't be given too much weight. You can use this information to design a better clustering algorithm.

10. Clustering review

03:02 - 03:34

Let's review. Clustering is an Unsupervised Machine Learning method that divides an unlabeled dataset into different categories. In order to perform clustering, we must first select relevant features of our dataset. Next, we select the number of clusters based on hypotheses about our data. Finally, we use the results of our clustering to solve our problems, whether it's defining new species, segmenting customers, or classifying movies into genres. There are a lot of diverse usages for clustering!

11. Let's practice!

03:34 - 03:39

Now that we understand clustering, let's practice!