MITOCW | MIT15_071S17_Session_6.2.07_300k

So how does clustering work?

The first step in clustering is to define the distance between two data points.

The most popular way to compute the distance is what's called Euclidean distance.

This is the standard way to compute distance that you might have seen before.

Suppose we have two data points-- I and J. The distance between the two points, which we'll call Dij, is equal to the square root of the difference between the two points in the first component, squared, plus the difference between the two points in the second component, squared, all the way up to the difference between the two points in the k-th component, squared, where k here is the number of attributes or independent variables.

Let's see how this works by looking at an example.

In our movie lens dataset, we have binary vectors for each movie, classifying that movie into genres.

The movie Toy Story is categorized as an animation, comedy, and children's movie.

So the data for Toy Story has a 1 in the spot for these three genres and a 0 everywhere else.

The movie Batman Forever is categorized as an action, adventure, comedy, and crime movie.

So Batman Forever has a 1 in the spot for these four genres and a 0 everywhere else.

So given these two data observations, let's compute the distance between them.

So the distance, d, would be equal to the square root of $(0-0)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2$, 1 et cetera.

This ends up being equal to the square root of 5.

In addition to Euclidean distance, there are many other popular distance metrics that could be used.

One is called Manhattan distance, where the distance is computed to be the sum of the absolute values instead of the sum of square.

Another is called maximum coordinate distance, where we only consider the measurement for which the data points deviate the most.

Another important distance that we have to calculate for clustering is the distance between clusters, when a cluster is a group of data points.

We just discussed how to compute the distance between two individual points, but how do we compute the distance between groups of points?

One way of doing this is by using what's called the minimum distance.

This defines the distance between clusters as the distance between the two data points in the clusters that are closest together.

For example, we would define the distance between the yellow and red clusters by computing the Euclidean distance between these two points.

The other points in the clusters could be really far away, but it doesn't matter if we use minimum distance.

The only thing we care about is how close together the closest points are.

Alternatively, we could use maximum distance.

This one computes the distance between the two clusters as the distance between the two points that are the farthest apart.

So for example, we would compute the distance between the yellow and red clusters by looking at these two points.

Here, it doesn't matter how close together the other points are.

All we care about is how close together the furthest points are.

The most common distance metric between clusters is called centroid distance.

And this is what we'll use.

It defines the distance between clusters by computing the centroid of the clusters.

The centroid is just the data point that takes the average of all data points in each component.

This takes all data points in each cluster into account and can be thought of as the middle data point.

In our example, the centroids between yellow and red are here, and we would compute the distance between the clusters by computing the Euclidean distance between those two points.

When we are computing distances, it's highly influenced by the scale of the variables.

As an example, suppose you're computing the distance between two data points, where one variable is the revenue of a company in thousands of dollars, and another is the age of the company in years.

The revenue variable would really dominate in the distance calculation.

The differences between the data points for revenue would be in the thousands.

Whereas the differences between the year variable would probably be less than 10.

To handle this, it's customary to normalize the data first.

We can normalize by subtracting the mean of the data and dividing by the standard deviation.

We'll see more of this in the homework.

In our movie data set, all of our genre variables are on the same scale.

So we don't have to worry about normalizing.

But if we wanted to add a variable, like box office revenue, we would need to normalize so that this variable didn't dominate all of the others.

Now that we've defined how we'll compute the distances, we'll talk about a specific clustering algorithm-- hierarchical clustering-- in the next video.