Unsupervised Machine Learning: Crash Course Statistics #37

Crash Course: Statistics
https://youtube.com/watch?v=IUn8k5zSI6g
https://nerdfighteria.info/v/IUn8k5zSI6g

Hi, I'm Adriene Hill, and welcome back to Crash Course Statistics.

In the last episode, we talked about using Machine Learning with data that already has categories that we want to predict. Like teaching a computer to tell whether an image contains a hotdog or not.

Or using health information to predict whether someone has diabetes. But sometimes we don't have labels. Sometimes we want to create labels that don't exist yet.

Like if we wanted to use test and homework grades to create 3 different groups of students in your Stats course. If you group similar students together, you can target each group with a specific review session that addresses its unique needs. Hopefully leading to better grades!

Because the groups don't already exist, we call this Unsupervised Machine Learning since we can't give our models feedback on whether they're right or not. There are no "True" categories to compare our groups with. Putting data into groups that don't already exist might seem kinda weird but today we'll explore two types of Clustering--the main type of Unsupervised Machine Learning: k-means and Hierarchical clustering.

And we'll see how creating new groups can actually help us a lot. INTRO Let's say you own a pizza restaurant. You've been collecting data on your customers' pizza eating habits.

Like how many pizzas a person orders a week. And the average number of toppings they get on their pizzas. You're rolling out a new coupon program and you want to create 3 groups of customers and make custom coupons to target their needs.

Maybe 2-for-1 five-topping medium pizzas. Or 20% off all plain cheese pizza. Or free pineapple topping!

So let's use k-means to create 3 customer groups. First, we plot our data: All we know right now is that we want 3 separate groups. So, what the k-means algorithm does is select 3 random points on your graph.

Usually these are data points from your set, but they don't have to be. Then, we treat these random points as the centers of our 3 groups. So we call them "centroids".

We assign each data point (the points in black) to the group of the centroid that it's closest to. This point here is closest to the Green center. So we'll assign it to the green group.

Once we assign each point to the group it's closest to, we now have three groups, or clusters. Now that each group has some members, we calculate the current centroid for each group. And now that we have the new centroids we'll repeat this process of assigning every point to the closest centroid and then recalculating the new centroids.

The computer will do this over and over again until the centroids "converge". And here, converge means that the centroids and groups stop changing, even as you keep repeating these steps . Once it converges, you have your 3 groups, or clusters.

We can then look at the clusters and decide which coupons to send. For example, this group doesn't order many pizzas each week but when they do, they order a LOT of toppings. So they might like the "Buy 3 toppings get 2 free" coupon.

Whereas this group, who orders a lot of simple pizzas, might like the "20% off Medium-2 topping-Pizzas" coupon. (This is probably also the pineapple group since really, there aren't that many things that pair well with pineapple and cheese.) If you were a scientist, you might want to look at the differences in health outcomes between the three pizza ordering groups. Like whether the group that orders a lot of pizza has higher cholesterol. You may even want to look at the data in 5 clusters instead of 3.

And k-means will help you do that. It will even allow you to create 5 clusters of Crash Course Viewers based on how many Raccoons they think they can fight off, and the number of Pieces of Pizza they claim to eat a week. This is actual survey data from you all.

A K-means clustering created these 5 groups. We can see that this green group is PRETTY confident that they could fight off a lot of raccoons. But 100 raccoons?

No. On the other hand, we also see the light blue group. They have perhaps more reasonable expectations about their raccoon fighting abilities, they also eat a lot of pizza each week.

Which makes me wonder…could they get the pizza delivery folks to help out if we go to war with the raccoons? Unlike the Supervised Machine Learning we looked at last time, you can't calculate the "accuracy" of your results because there's no true groups or labels to compare. However, we're not totally lost.

There's one method called the silhouette score can help us determine how well fit our clusters are even without existing labels. Roughly speaking, the silhouette score measures cluster "cohesion and separation" which is just a fancy way of saying that the data points in that cluster are close to each other, but far away from points in other clusters. Here's an example of clusters that have HIGH silhouette scores.

And here's an example of clusters that have LOW silhouette scores. In an ideal world, we prefer HIGH silhouette scores, because that means that there are clear differences between the groups. For example, if you clustered data from lollipops and Filet Mignon based on sugar, fat, and protein content the two groups would be VERY far apart from each other, with very little overlap--leading to high silhouette scores.

But if you clustered data from Filet Mignon and a New York Strip steak, the data would probably have lower silhouette scores, because the two groups would be closer together - there'd probably be more overlap. Putting data into groups is useful, but sometimes, we want to know more about the structure of our clusters. Like whether there are subgroups--or subclusters.

Like in real life when we could look at two groups: people who eat meat and those who don't. The differences between the groups' health or beliefs might be interesting, but we also know that people who eat meat could be broken up into even smaller groups like people who do and don't eat red meat. These subgroups can be pretty interesting too.

A different type of clustering called Hierarchical Clustering allows you to look at the hierarchical structure of these groups and subgroups. For example, look at these ADORABLE dogs. We could use hierarchical clustering to cluster these dogs into groups.

First, each dog starts off as its own group. Then, we start merging clusters together based on how similar they are. For example, we'll put these two dogs together to form one cluster, and these two dogs together to form another.

Each of these clusters--we could call this one "Retrievers" and this one "Terriers", is made up of smaller clusters. Now that we have 2 clusters, we can merge them together, so that all the dogs are in

one cluster. Again, this cluster is made up of a bunch of sub clusters which are themselves made up of even smaller sub clusters.

It's turtles I mean clusters all the way down. This graph of how the clusters are related to each other is called a dendrogram. The further up the dendrogram that two clusters join, the less similar they are.

Golden and Curly Coated Retrievers connect lower down than Golden Retrievers and Cairn Terriers. One compelling application of hierarchical clustering is to look for subgroups of people with Autism Spectrum Disorder--or ASD. Previously, disorders like Autism, Aspergers and Childhood Disintegrative Disorder (CDD) were considered separate diagnoses, even though they share some common traits.

But, in the latest version of the Diagnostic and Statistical Manual of Mental Disorders--or DSM - these disorders are now classified as a single disorder that has various levels of severity, hence the Spectrum part of Autism Spectrum Disorder. ASD now applies to a large range of traits. Since ASD covers such a large range, it can be useful to create clusters of similar people in order to better understand Autism and provide more targeted and effective treatments.

Not everyone with an ASD diagnosis is going to benefit from the same kinds and intensities of therapy. A group at Chapman University set out to look more closely at groups of people with ASD. They started with 16 profiles representing different groups of people with an ASD diagnosis.

Each profile has a score between 0 and 1 on 8 different developmental domain. Low scores in one of these domains means it might need improvement. Unlike our pizza example which had only 2 measurements--# of pizza toppings and # of pizzas ordered per week--this time we have 8 measurements.

This can make it tough to visually represent the distance between clusters. But the ideas are the same. Just like two points can be close together in 1 or 2 dimensions, they can be close together in 8 dimensions.

When the researchers looked at the 16 profiles, they grouped them together based on their 8 developmental domain scores. In this case, we take all 16 profiles and put each one in their own "cluster", so we have 16 clusters, each with one profile in them. Then, we start combining clusters that are close together.

And then we combine those , and we keep going until every profile is in one big cluster. Here's the dendrogram. We can see that there are 5 major clusters, each made up of smaller clusters.

The research team used radar graphs, which look like this, to display each cluster's 8 domain scores on a circle. Low scores are near the center, high scores near the edge of the circle. This main cluster, which they called Cluster E, has scores consistent with someone who is considered high functioning.

Before the change to the DSM, individuals in the cluster might have been diagnosed with Asperger's. The Radar graph here shows the scores for the 6 original data points that were put in Cluster E. While there are some small differences, we can see that overall the patterns look similar.

So Cluster E might benefit from a less intense therapy plan, while other Clusters with lower scores--like Cluster D--may benefit from more intensive therapy. Creating profiles of similar cases might allow care providers to create more effective, targeted therapies that can more efficiently help people with an ASD diagnosis. If an individual's insurance only covers say 7 hours of therapy a week, we want to make sure it's as effective as possible.

It can also help researchers and therapists determine why some people respond well to treatments, and others don't. The type of hierarchical clustering that we've been doing so far is called Agglomerative, or bottom-up clustering. That's because all the data points start off as their own cluster, and are merged together until there's only one.

Often, we don't have structured groups as a part of our data, but still want to create profiles of people or data points that are similar. Unsupervised Machine Learning can do that. It allows us to use things that we've observed--like the tiny stature of Terriers, or raccoon-fighting confidence --and create groups of dogs, or people that are similar to each other.

While we don't always want categorize people, putting them into groups can help give them better deals on pizza, or better suggestions for books or even better medical interventions. And for the record, I am always happy to help moderately confident raccoon fighting pizza eaters fight raccoons. Just call me.

Thanks for watching. I'll see you next time.