UNIT 4: Unsupervised Learning

Tabaré Pérez

May 6, 2020

UNIT OVERVIEW

So in this unit, we will cover unsupervised learning.

At first, the idea of unsupervised learning sounds totally counterintuitive. Because now we don't have any labels. How do we know the machine actually learns anything?

In contrast to the supervised learning that we covered so far in the class, where we have really clear objectives that we are trying to optimize, now we will kind of re-think this question, of what does it mean to find structure in the data.

And I should say that in daily life, we use a lot of type of unsupervised learning. For instance, if you are looking at stories that you read everyday at Google News, you would see that a machine can automatically cluster different stories written about the same event, so it does introduce to us a structure over the whole stream of the news.

So we will start by looking at a hard assignment clustering algorithm where we assume that a point, or a news story, can just belong to a single cluster. We'll provide the algorithm with an idea of similarity between points, and ask it to find the best possible arrangement to cluster these points. Next, we will move to more exciting representation of the data structure.

Because many times, it's actually pretty hard for us to say at this point, that this news story really belongs to this class.

Maybe it belongs to several clusters, depending on how we structure the data. So a much richer way to think about structure is to think about it in probabilistic terms, by just looking at different classes of data, different components of data or looking at the likelihood a certain point was generated by that component.

And we will introduce generative model, which provide you the probability distribution over the points.

We will look at multinomials, Gaussian and mixture of Gaussians.

And by the end of this unit, we will get to a very powerful unsupervised learning algorithm for uncovering then, the line structure, which is called the **E-M**.

You will have a chance to practice what you learn in this unit in your project, which has a lot of practical application on Netflix recommendation system, using both clustering and the **E-M**.