Lecture 13 - 5: Clustering Definition

Tabaré Pérez

May 6, 2020

So now we are ready to start thinking about clustering in more formal terms.

And the first way of thinking about clustering is thinking about it as partitioning.

So what we assume that the **INPUT** to the clustering will be just a set of feature vectors, and you would also be given capital K and capital K is the number of clusters.

$$S_n = \{x^{(i)} | i = 1 \dots n\}, K \tag{1}$$

In our case, for instance, in the cartoon example with image quantization, it would be the number of colors. This is the input.

The **OUTPUT** to this algorithm would be partitions:

$$C_1 \dots C_K \tag{2}$$

So each partition would do record the index of the elements that belong to this cluster.

So for instance, what we can say more formally what does it mean to have partition is that, if you make a union of all this partitions, you will get all the elements of your training set.

So in this case, if we take the union of all these different clusters, we will get all the elements in our original set from 1 to n.

$$C_1 \dots C_K; \bigcup C_j = \{1 \dots n\} \tag{3}$$

And I want to emphasize that we are not remembering the element itself of the element x, but the index of the element.

So you would cover when you look at the union all the indexes of the elements in your original set.

And another conditions that you will have is that, if we look at some C_i and C_j for different i and j, their intersection will be empty:

$$C_i \bigcap C_j = \phi, (i \neq j) \tag{4}$$

So in other words, what we're going to be looking at is hard clustering, where every element just belongs to one partition.

So this is one way to think of clustering is just grouping over the elements.

There is another important view, which we will use today during the lecture is to think about clustering in terms of their representatives.

We can think about clustering as selecting representatives:

Representatives:

$$z^{(1)} \dots z^{(K)} \tag{5}$$

So representatives in this case, will be vectors which would represent every single partitioning.

So if we go back to Google News example, we can think that the cluster is actually the stories that I selected to represent this particular set of news articles.

Or in the case of image quantization, you can think that the representative would be the color that I selected to represent all these pixels that got collapsed into a single color.

And we will see today what is the connection between the two views.

Clearly they are connected:

- If you know what is your partitioning, you may be able to guess who is the representative.
- If you know who is a representative, you may be able to guess who is a constituent.

You will see how you can actually unify these two views together.

So at this point, we completed the discussion about clustering definition, and we've seen some examples. So the next part for us is actually to start talking about the clustering course.