**How scaling is useful in clustering analysis**

We perform clustering analysis to categorize the batteries which are similar. Batteries which are similar will possess common characteristics i.e. in other words, the values of the different attributes/parameter/features would have nearly same values.

In Machine Learning, the similarity is measured in terms of distance.

So when we have very large number of batteries, it is very useful to find batteries which are similar, thus it helps us to divide batteries into different segments.

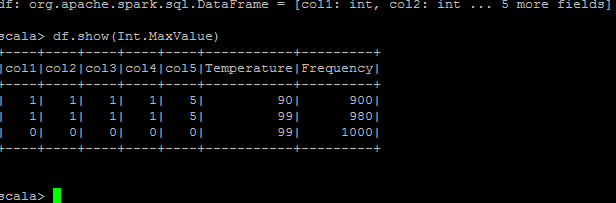
We do this segmentation or categorization based on clustering analysis. So inside a cluster, we expect to have batteries which are similar i.e. the batteries whose values of the different attributes/parameter/features would have nearly same values.

In many scenarios, it becomes very important to scale the features/ attributes so that the different attributes/features/parameters of a battery have a proportional contribution in distance measurement. Because cluster are formed based on distance between two batteries.

Let us suppose we have 3 batteries, each battery have 7 features/attributes. The first 4 attributes are can have only binary values i.e. 0 or 1 the 5 attributes have a range of values between 0 and 10 and the last two attributes let’s say temperature and frequency have values ranging 0 to 100 and 100 to 2000 respectively. Our goal is two form two cluster and group the two similar batteries in one cluster and 3rd one in other cluster.

Original Dataset of 3 batteries with attributes/features values

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Battery | Col1 | Col2 | Col3 | Col 4 | Col 5 | Temperature | Frequency |
| 1 | 1 | 1 | 1 | 1 | 5 | 90 | 900 |
| 2 | 1 | 1 | 1 | 1 | 5 | 99 | 980 |
| 3 | 0 | 0 | 0 | 0 | 0 | 100 | 1000 |



We can easily infer from above table that battery 1 and 2 are similar and should be in one cluster and battery 3 in other cluster (since we are having two clusters)

We find distance between all pair of batteries and on calculation we find distance between battery 2 and 3 is minimum and thus we end up grouping battery 2 and 3 in one cluster and battery 1 in other.

Distance is calculated using Euclidean distance:

<http://mathonline.wikidot.com/the-distance-between-two-vectors>

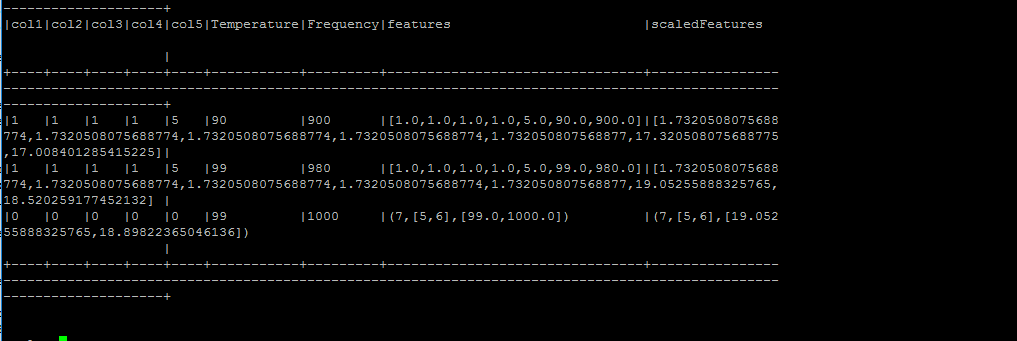
For calculation refer the below link:

Incorrect Cluster

We know this is wrong grouping as battery 1 and 2 are more similar. So what went wrong?

It is because temperature and frequency can take very large values as compared to first 5 attributes, they dominate the first 5 attributes and thus the contribution of first 5 attributes in distance calculation becomes negligible as compared to contribution of temperature and frequency attribute in distance calculation. Thus we ended up with wrong clusters.

Let us see the effect of scaling now. On scaling the attributes our original dataset is transformed as below



Transformed Data using Standard Scaler

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Battery | Col1 | Col2 | Col3 | Col4 | Col5 | Temperature | Frequency |
| 1 | 1.7 | 1.7 | 1.7 | 1.7 | 1.7 | 17.3 | 17.0 |
| 2 | 1.7 | 1.7 | 1.7 | 1.7 | 1.7 | 19.0 | 18.5 |
| 3 | 0 | 0 | 0 | 0 | 0 | 18.1 | 18.9 |

In this case with transformed dataset, we will have lesser distance between battery 1 and two and thus cluster will form with battery one and two. We can see clearly that scaling helps to make all attributes proportional such that all factors contributes in distance calculation and thus form a correct cluster.

Correct Cluster