k-means Clustering

Jan Gaura

2019-03-13

k-*means Clustering*

When we implemented classification using etalons, we had to manualy assign a class to each object. We may say that this approach is a small example of supervised learning.

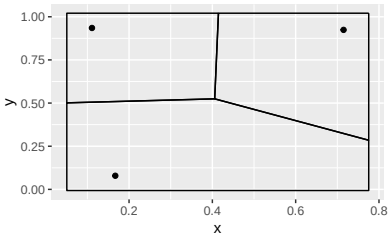
Sometimes, we would like not to assign anything manualy and still get the classification right. One such a method is called *k*-means clustering. This method is able to cluster (basicaly separate) input data in k different partions (clusters). Nice feature is that it is able to do so automatically without our input. As you may already guessed, the k in its name stands for number of clusters (classes) in the input data. In our case the $k = 3$.

Algorithm[1]

[1] https://en.wikipedia.org/wiki/K-means_clustering

0. Initialize k centroids $(m_1, \ldots, m_k)$ to randomly selected points from the input.
1. Compute Euclidean distance from each centroid to all input data points.
2. Assign each input data to the closest centroid.
3. Update position of the centroid by calculating the mean position of the assigned points.
4. Repeat from step 1 until centroids do not move very much (distance is less that give threshold, or no reassignment of data points occurs).

The algorithm basically creates a voronoi diagram for centroids.



It's a bit unfortunate that the algorithm does not always converge. So you have to run it once again to obtain a desired output.

Also, the output depends on initialization, you can experiment with the initial positions of centroids to see what happens.

Centroids provided by *k*-means algorithm are then used in classification in the same way as were etalons. The only difference is that we do not know which class of object is represented by which centroid. Classification than works in the way that class names are simply numbers.