K means accelerator

## Introduction

### The K means algorithm

The K means algorithm is an iterative algorithm which divides a given data vector to K different clusters (K is a natural number). Each cluster will be characterized by its “center of mass”, what will be referred in this paper as centroid.

#### The algorithm steps

For a simpler explanation, we can assume K is a constant predefined natural value. First, will define some symbols:

-the cluster number "*i*" centroid

– the group of points in cluster number "*i*"

##### Initiation step

The first step in the algorithm is to randomly choose centroids for the K clusters. The “time” (“*t*”) for the initialization step will be defined as zero.

##### Classification step

In each iteration(time) of the algorithm we first assign each points from the input vector to a cluster based on the “distance” from the point to the clusters centroid. A point will be assigned to cluster number “i” if the metrical distance between it and the cluster’s centroid is the minimum between the distances from the point to all others cluster’s centroids. To simplify:

\*In case of the distance from two different clusters is the same and is the minimum found, the chosen cluster is the one with the lowest index.

##### Centroids update step

After the classification step, the centroids of each cluster are updated to be mean of all points which belong to it in end of iteration(time) *t*. This is done by verifying if a cluster is empty(in this case the centroid is not changed) and then calculating the mean of all the clusters points:

##### Converge check step

If the centroids of the next iteration calculated in the step above are all equal to the current centroids, then the algorithm comes to an end. Else, the iteration number(time) is increased by one and a new iteration begins with the assigning step.

Note : the k means algorithm assures convergence to a local minimum.

#### Choosing K

Usually the optimal K is not known before the beginning of the algorithm. Therefore, a an error parameter can be defined to help choosing K. The most commonly known error parameter is the clustering error which is defined by:

In this formula, the elements are:

As K increases, the error decreases. For example, if K is as the number of pints in the input vector, the error will be zero. This because it cluster will have just one point which will also be its centroid, but in this case no new information was added by the algorithm.

One suggested method of choosing a natural K so the clustering error is minimized is by gradually increasing K and calculating for each increasement. The process ends when the error reaches a value so that , where is a predefined threshold.