K-Means

Algorithm description

Approach: Cluster data by separate samples to n groups of equal variance. The objective is to minimise inertia (Within cluster sum of squares criterion).

Inertia criterion equation:

Algorithm

1. Choose centroids
   loop until number of iterations is met

   Loop body:
2. Assign each sample to its nearest centroid (label update step)

3. Recompute centroids for the next iteration by taking mean of alll samples assigned to each centroid.

   If new centroids - old centroids is smaller than a threshold, break out of the loop. If the inertia calculation from new centroids isn't smaller than current minimum don't reassign the centroids.

Scalability: Scales well.

Disadvantages:

Clusters assumed to be convex and isotropic (Separable and of equal variance).

That is to say, the algorithm will not perform well on clusters with irregular shapes/elongated clusters

Inertia is not a normalised metric. In very high-dimensional spaces, can suffer from the curse of dimensionality

Coded algorithm with sample dataset