SPHERICAL K-MEANS CLUSTERING

Akshai James (194102301)

15 Mar2020

1 Theory

Spherical k-means is an unsupervised clustering algorithm where the lengths of all vectors being compared are normalized to 1,so that they differ in direction but not in magnitude.Clustering can then be carried out more efficently by measuring the angles between the vectors.
Spherical k-means is preferred to standard k-means:

- When the magnitude of the vectors is irrelevant in terms of what the data represents.
- When the magnitude of the vectors is not particularly important in terms of what the data represents and the vectors have a large number of dimensions,because spherical k-means is a more efficent learning technique.

The k-means method of clustering minimizes the sum of squared distances between cluster centres and cluster members.If the entities to be clustered are projected on the unit sphere,then a natural measure of dispersion is the sum of squared chord distances seperating the entities from their cluster centers;k-means clustering with this measure of dispersion is spherical k-means.
For high- dimensional data such as text documents,this method has been shown to be a superior method.The

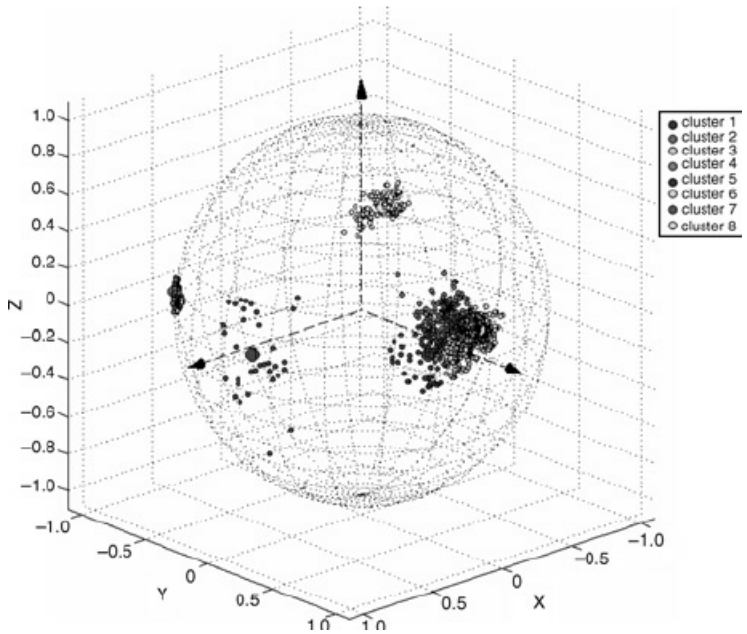

Figure 1: Spherical k-means clustering

implication is that the direction of a document vector is more important than the magnitude.This leads to a unit vector representation,i.e.,each document vector is normalized to be of unit length.Spherical k-means algorithm involves mainly two steps.In the first step,k data points are randomly choosen as mean centres and it is called as the seeding step.In the second step,through consecutive iterations,make k clusters by measuring the angle between data points to mean centres.

1.1 Algorithm

Consider the unit vecors $\{x_i\}$;i=1,2,....,N to be clustered into K groups,where each group is represented by unit vector $\{\mu_j\}$;j=1,2,...K. Steps:

- Randomly initialise the labels to each data points l_i
- $V_j(t) = \frac{\sum_{i=1}^{n} x_i \delta_k[l_i(t-1)-j]}{\sum_{i=1}^{n} \delta_k[l_i(t-1)-j]}$
- $\hat{\mu_j}(t) = \frac{V_j(t)}{\|V_j(t)\|}$
- $l_i(t) = \arg\max\{\hat{x_i}^T \hat{\mu_j}(t)\}$ for j =1,2...K
- $\min\{\hat{\mu_j}(t)^T \mu_j(\hat{t-1})\} \geq 0.99$ for j=1,2,...K

1.2 Description

Step 1:We randomly initialising labels to all the data points.No of distinct labels will be equal to no of clusters we needed.

Step 2:Here we are updating the mean value of each cluster,using the given equation above.The numerator $\sum_{i=1}^{n} x_i \delta_k[l_i(t-1)-j]$ actually doing the summation of all data points which are belong to one cluster.When the label value become equal to cluster value,the delta function will become one.So it adds all data points which has same cluster value.In each iteration,cluster mean changes,until termination condition is reached.The denominator $\sum_{i=1}^{n} \delta_k[l_i(t-1)-j]$ counts the no of data points belongs to corresponding clusters.

Step 3:The means corresponding to each cluster find out in Step 2 will have length greater than one.So to make them of unit magnitude,we are normalising the cluster means.

Step 4:Relabeling the ith data point.Here we are taking dot product of ith data point with all cluster means.So the new label of ith data point will be that cluster index which gives maximum dot product.

Step 5:Here termination condition is checked.If This criteria is satisfied,termination will happen.Otherwise,iterations will continue.