K-means Clustering algorithm

Charan Shetty

May 14,2014

Abstract: K means clustering is a unsupervised learning algorithm for classification of data items.Cluster analysis will have data which has no labels associated with it.The algorithm tries to finds a structure or clusters in the data points and hence exploratory in nature.We provide a brief overview of k means clustering,challenges in implementation and an illustration on how k means can be used to cluster terms.

Introduction: K means algorithm can be stated as follows:Given n data points, find k groups on a measure of similarity between the data points such that the similarity is high for the data points within the group while it is low between the data points of different groups[6].

Algorithm:

```
input: -input data set $(x^{(1)}, x^{(2)}........, x^{(m)})$
        -k (number of clusters)
Randomly initialize cluster centroids as it is an iterative algorithm
Loop{
 for $l$ =1 to $m$
 $c^{(i)}$=index (from 1 to $K$) of cluster centroid closest to $x^{(i)}$
 for $k$=1 to K
     $\mu_k$=avg mean of data points assigned to $k$ cluster
}
```

Choosing number of clusters(k): Finding the right value of k is the most important part of algorithm and there is no agreed upon solution[8].It is ambiguous to get clusters in the data.As shown in Figure 1.(clusters of normal distributions were taken and plotted [6]) It is quite unclear to say exactly how many clusters are there.It can be either 2 or 3 or 4.While in Figure 2 its difficult find number of clusters.

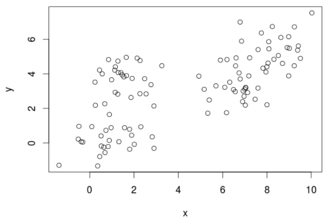


Figure 1: Clearly identifiable clusters

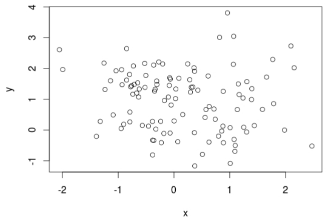


Figure 2: Clusters not clearly identifiable

Elbow method can be used to find the number of clusters[8] but it doesn't give correct value in all cases.So for the above plot "Figure 1",Using this method a plot on cost function versus number of clusters is drawn.For k¿3 there is no significant change or decrease in the cost function(error),while for k=2,3 there is a sudden change in slope(like the elbow of the arm)of curve implying these are the better choices for k.Now as mentioned it is not guaranteed to give correct results. So for elbow method plot for "Figure 1" there is no good k value as there is no k value in the plot where there is a sudden change in slope as it is more like negative exponential plot(Figure 4).The plots were implemented by taking clusters of normal distributions and plotted in R[6]

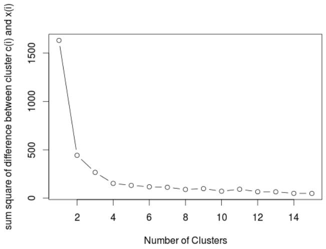


Figure 3: Plot of elbow method for figure 1

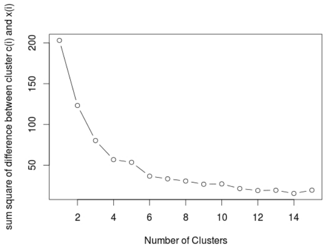


Figure 4: plot of elbow method of figure 2

Initialsing cluster centroid: Now that we have number of clusters to be looked at we have to position these k cluster centroids.We can randomly intialize the cluster centroids,There are many ways to do this,One better option would be that ,
for k< numberof datapoints Randomly pick k data points
Label them as μ_1,μ_2,......μ_k
So if these centroids belong to different clusters we can reach the local optima which can be the global optima.But it can so happen that these

randomly picked data points belong to same cluster.In that case it can get stuck in bad local optima.To solve this problem we can initialize randomly multiple times the cluster centroids and calculate the cost function.Choose the clustering instance that gave a lower cost function.

Applying kmeans algorithm to term document matrix:
K-means algorithm can be applied to term document matrix to find the possible clusters in the matrix. To find cluster of related terms(terms that co occur) We considered 4 documents related to 2 mutually exclusive topics ”cricket rules” and ”birds” taken from articles in wikipedia[7] and checked if we can get the 2 clusters from the corpus. The steps for implementing it were following
1.The punctuations in the text were removed
2.All letters were converted to lowercase,extra whitespaces and numbers were removed
3.The Stopwords from the document were removed.The stop word set was as per the frequent words in english[3].
4.Term document matrix of the corpus was created.
5.This matrix had frequency of terms repeated in every document.
6.K means algorithm was run on this matrix for 2 clusters[4]. As shown in the Figure 5 clusters were created with datapoints ”1” belong to cluster 1 and datapoints ”2” belong to cluster 2.x and y co ordinates ”dc1” and ”dc2” being the Discriminant coordinates/canonical variates[5]

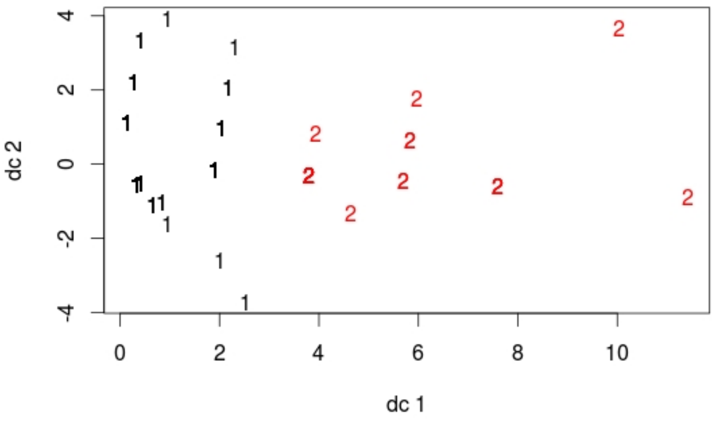
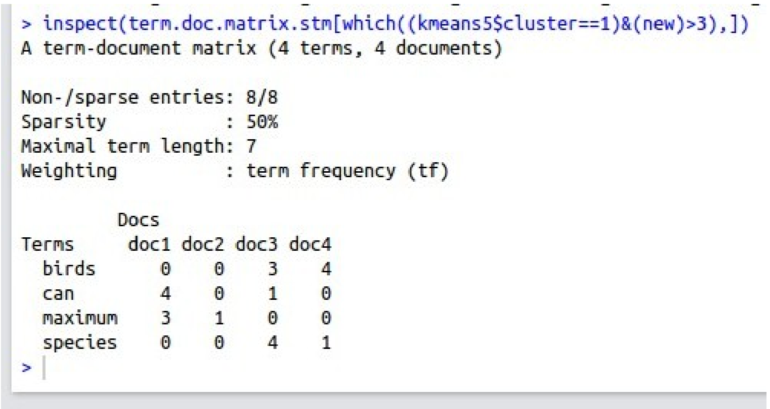


Figure 5: Showing 2 clusters labelled ”1” and ”2”

The below figures(Figure 6 and figure 7) shows the most frequent words in the 2 clusters cluster 1 and cluster 2.From this figure it is clear that the k means cluster was able to classify the two topics in the dataset as 2 clusters and was able to classify correctly.That is terms related to cricket belongs to cluster 2 while terms belonging to ”birds” belongs to cluster 1.Since it being a smaller corpus clustering is not very accurate,however with larger corpus it is not a problem.

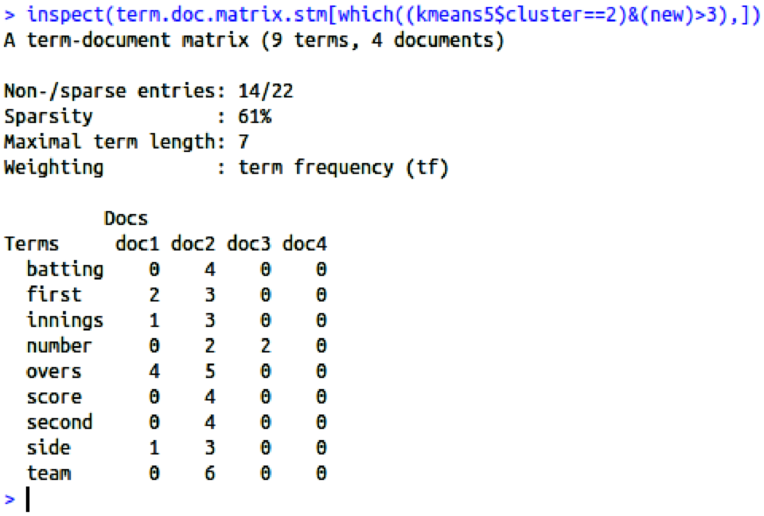


```
> inspect(term.doc.matrix.stm[which((kmeans5$cluster==1)&(new)>3),])
A term-document matrix (4 terms, 4 documents)

Non-/sparse entries: 8/8
Sparsity           : 50%
Maximal term length: 7
Weighting          : term frequency (tf)

         Docs
Terms     doc1 doc2 doc3 doc4
  birds      0    0    3    4
  can        4    0    1    0
  maximum    3    1    0    0
  species    0    0    4    1
> 
```

Figure 6: frequent words in cluster ”1”



```
> inspect(term.doc.matrix.stm[which((kmeans5$cluster==2)&(new)>3),])
A term-document matrix (9 terms, 4 documents)

Non-/sparse entries: 14/22
Sparsity           : 61%
Maximal term length: 7
Weighting          : term frequency (tf)

         Docs
Terms     doc1 doc2 doc3 doc4
  batting    0    4    0    0
  first      2    3    0    0
  innings    1    3    0    0
  number     0    2    2    0
  overs      4    5    0    0
  score      0    4    0    0
  second     0    4    0    0
  side       1    3    0    0
  team       0    6    0    0
> 
```

Figure 7: Frequent words in cluster 2

Conclusion: Once we are able to get clusters we can get an idea on what each cluster is about.This can be found out from the dominant terms in the cluster.There are many variants of K means cluster and this is the most widely used clustering algorithm[1] Inspite of many clustering algorithms it is still a difficult problem mainly because of ambiguity in defining a cluster and with the similarity measure.

References:

[1] Anil K. Jain (2009) Data Clustering: 50 Years Beyond K-Means.Michigan State University
[2] http://en.wikipedia.org/wiki/K-means_clustering
[3] http://www.r-bloggers.com/clustering-search-keywords-using-k-means-clustering
[4] http://randyzwitch.com/rsitecatalyst-k-means-clustering
[5] Seber, G. A. F. (1984). Multivariate Observations. New York: Wiley.
[6] https://github.com/charanshetty/kmeans
[7] http://en.wikipedia.org/wiki/Bird
[8]http://www.cs.princeton.edu/courses/archive/spr07/cos424/scribe_notes/0306.pdf