PC Project Report

Liron Mizrahi

October 2016

1 Introduction

K-means is one of the simplest unsupervised learning algorithms to solve a clustering problem. Clustering is the process of partitioning a group of data points into a small number of clusters. In general, a cluster is defined as a set of similar objects. The similarity in a given set may vary according to data, therefore a clustering algorithm that finds an optimization in one set of data may not find an optimization in another set of data.

There are many variations of the k-means clustering algorithm, however we have looked at 2 of them, namely the normal k-means clustering and online k-means clustering.

2 Terminology and definitions

For this report we will define some terms and variables which will be used:
There is a dataset $S \in R^m$.
$\underline{x}$ is a data point in S.
k is the number of clusters.
$\underline{\mu_j}$ is a cluster centre.
$\underline{d(\underline{x}, \underline{\mu_j})}$ is the distance from $\underline{x}$ to $\underline{\mu_j}$.
N_j is the number of points in cluster j.
The sum of squared error is given by:

$$\sum_j \sum_{\underline{x}} ||\underline{x} - \underline{\mu_j}||^2$$

where $|| \cdot ||$ is the Euclidean distance.
η is the learning rate where $\eta \in \big[0, 1\big]$. The learning rate is used for the online

k-means. It is used to increase or decrease the shift of the centres for each iteration.

3 Background Theory

As stated, k-means is a solution technique to solve a clustering problem. Clustering is the task of grouping a set of objects in such a way that objects in the same group have more similarities to each other than to those in other groups. Clustering is not just one specific algorithm, but it is a problem to be solved. The idea of a cluster cannot be precisely defined, as such different variations of the solution can be derived depending on the idea used of what constitutes a cluster and the method used to find them. This is one of the reasons why there are so many different clustering algorithms. Different cluster models are used by different people depending on their needs, and for each of these cluster models again different algorithms can be given. The cluster model used by k-means is called a centroid model, where each cluster is represented by a single mean vector.

4 The Clustering Problem

In centroid clustering, clusters are represented by a vector, which may not necessarily be a member of the dataset. When the number of clusters is fixed to k, k-means clustering becomes an optimization problem,i.e. find the k cluster centers and assign the points to the nearest cluster center, such that the squared distances from the cluster are minimized, i.e. the sum squared error is minimized. This problem is known to be NP-Hard.

5 k-means vs Online k-means

K-means and online k-means are very similar in that they accomplish the same task. However, there are some big differences. K-means loops through through all the data given and updates the centres after every point has been assigned to their respective closest centre. Each data point can be looked at individually as calculations for the closest centre for each point are not dependent on other points. Each centre is also independent of each other after the points have been assigned to them. This implies that k-means can benefit highly from parallelization. Chunks of data points as well as cluster centres can be processed concurrently.

On the other hand, online k-means is used to deal with continuous streams of data, unlike k-means where all the data is known beforehand. Online k-means looks at one data point at a time and finds the closest centre. Once the centre is found, it is then updated immediately. The problem with this is that calculations are now dependent on prior calculations. As seen above in the online k-means algorithm, the learning rate, η, is decreased after each iteration, meaning that the calculations for the centres will change depending on the order the data is given. and therefore online k-means is not suitable at all for parallelization.

6 Theoretical Analysis

The computational complexity of finding the optimal solution to the k-means clustering problem is known to be NP-hard for a general number of clusters k (even for 2 clusters) ans as such a heuristic algorithm must be used such as Lloyd's k-means algorithm shown above. The running time of Lloyd's algorithm is often given as $O(nkmi)$, where n is the number of vectors in $\mathbb{R}^m$ space, k the number of clusters and i the number of iterations needed until convergence.

7 Algorithms

Shown below is the general pseudo-code for the serial k-means clustering algorithm and the serial Online k-means clustering algorithm:

Given dataset S in $\mathbb{R}^m$;
Choose k, which is number of clusters;
Choose the k cluster centres(Randomly): $\underline{\mu_1}, \underline{\mu_2}, \ldots, \underline{\mu_k}$;
while *stopping condition does not hold* **do**
  for *each* $\underline{x} \in S$ **do**
    Find $min\{d(\underline{x}, \underline{\mu_1}), d(\underline{x}, \underline{\mu_2}), \ldots, d(\underline{x}, \underline{\mu_k})\}$;
    Say $d(\underline{x}, \underline{\mu_j})$ is the minimum;
    Assign $\underline{x}$ to cluster j;
  end
  for *each cluster centre* $\underline{\mu_j}$ **do**
    move cluster centre to the mean of the points in its cluster;

$$\underline{\mu_j} \leftarrow \frac{1}{N_j} \sum_{\underline{x}} \underline{x}$$

    where N_j is the number of points in cluster j.
  end
end

Algorithm 1: k-means clustering algorithm (This heuristic algorithm is also known as Lloyd's k-means algorithm

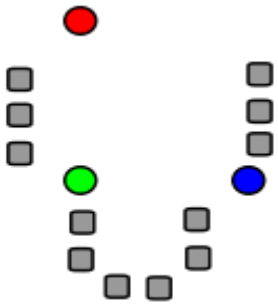
Given dataset S in $\mathbb{R}^m$;
Choose k, which is number of clusters;
Choose the k cluster centres(Randomly): $\underline{\mu_1}, \underline{\mu_2}, \ldots, \underline{\mu_k}$;
while *stopping condition does not hold* **do**
  for *each* $\underline{x} \in S$ **do**
    Find $min\{d(\underline{x}, \underline{\mu_1}), d(\underline{x}, \underline{\mu_2}), \ldots, d(\underline{x}, \underline{\mu_k})\}$;
    Say $d(\underline{x}, \underline{\mu_j})$ is the minimum;
    Update $\underline{\mu_j}$;

$$\underline{\mu_j} \leftarrow \underline{\mu_j} + \eta(\underline{x} - \underline{\mu_j})$$

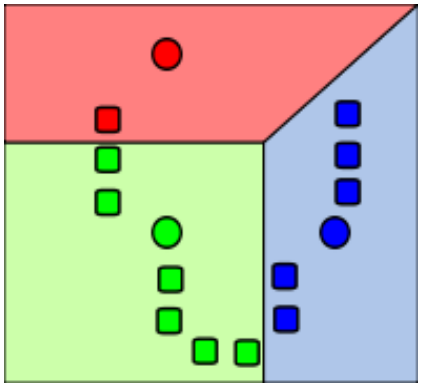
    where η is the learning rate.
  end
  Decrease η (This is an extra stopping condition).
end

Algorithm 2: Online k-means clustering algorithm

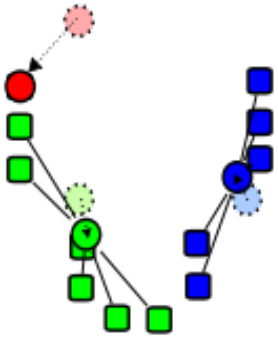
The stopping conditions may vary by time or by iterations. However, we have chosen the stopping condition to be bounded by the number of iterations.



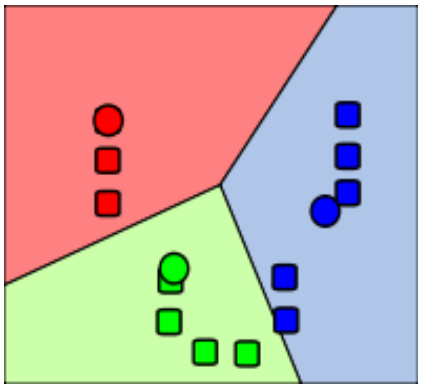
(a) k initial centres are randomly generated (in this case k=3).



(b) k clusters are created by associating every data point with the nearest centre.



(c) The centre of each of the k clusters becomes the new mean of the points.



(d) Steps (b) and (c) are repeated until convergence has been reached.

Figure 1: Example demonstrating how k-means works