Map Reduce for k-Means

Algorithm has several steps:

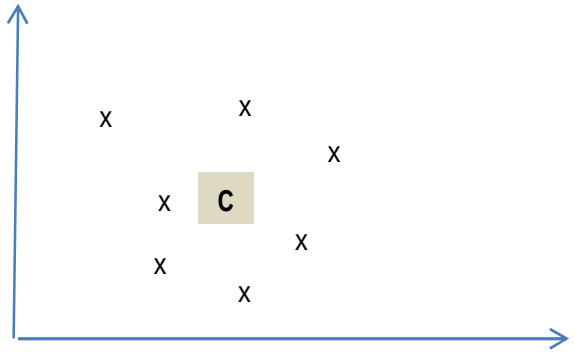
1. Initialize centroid guess

Iterate:

2. Assign points to closest guessed centroid
3. Form new guess by calculating centroid of assigned points

Centroid Calculation

How do we calculate the centroid of a collection of points in a vector space?



The centroid of a set of points (vectors) is their (vector) average. Sum of points divided by number of points.

Assume:

1. we've got a guess for the centroids.
2. the mappers have the centroid guess

Map Reducing k-Means

-Seems like "sum of points" is a good thing to focus on for the sufficient statistic from the mapper (following statistical query model).

-First the mapper has to decide which points go with which centroids.

-The mapper algorithm is something like:

Initialize k accumulators to 0 vector. (1 for each centroid).

For each point:

1. Assign to the closest centroid
2. Add the point (vector addition) to the corresponding accumulator.

When finished emit partial sums AND number of points constituting each partial sum.