ROADMAP

Kmeans

Mapreduce implementation

Demonstration

Results and statistics

KMEANS

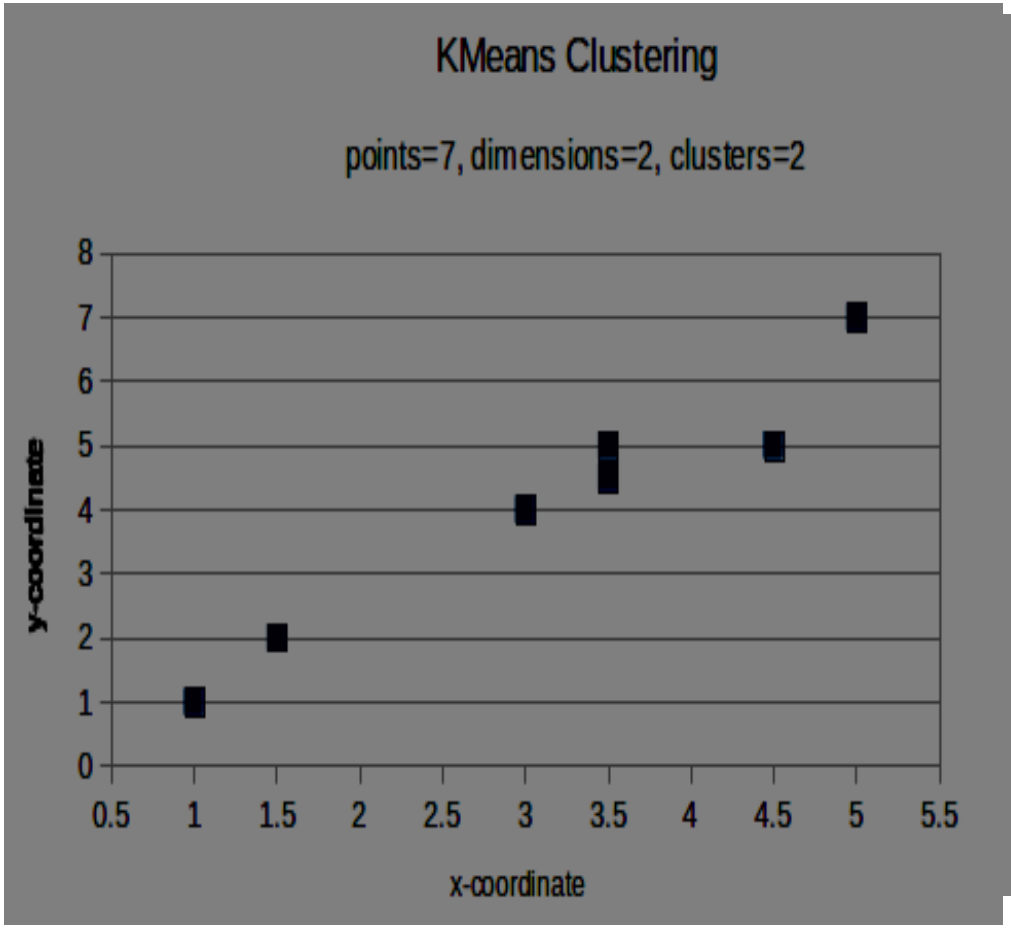
Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \leq n$ and $k>0$) $S = \{S_1, S_2, \ldots, S_k\}$ based on features/attributes of data points so as to minimize the within-cluster sum of squares (WCSS):

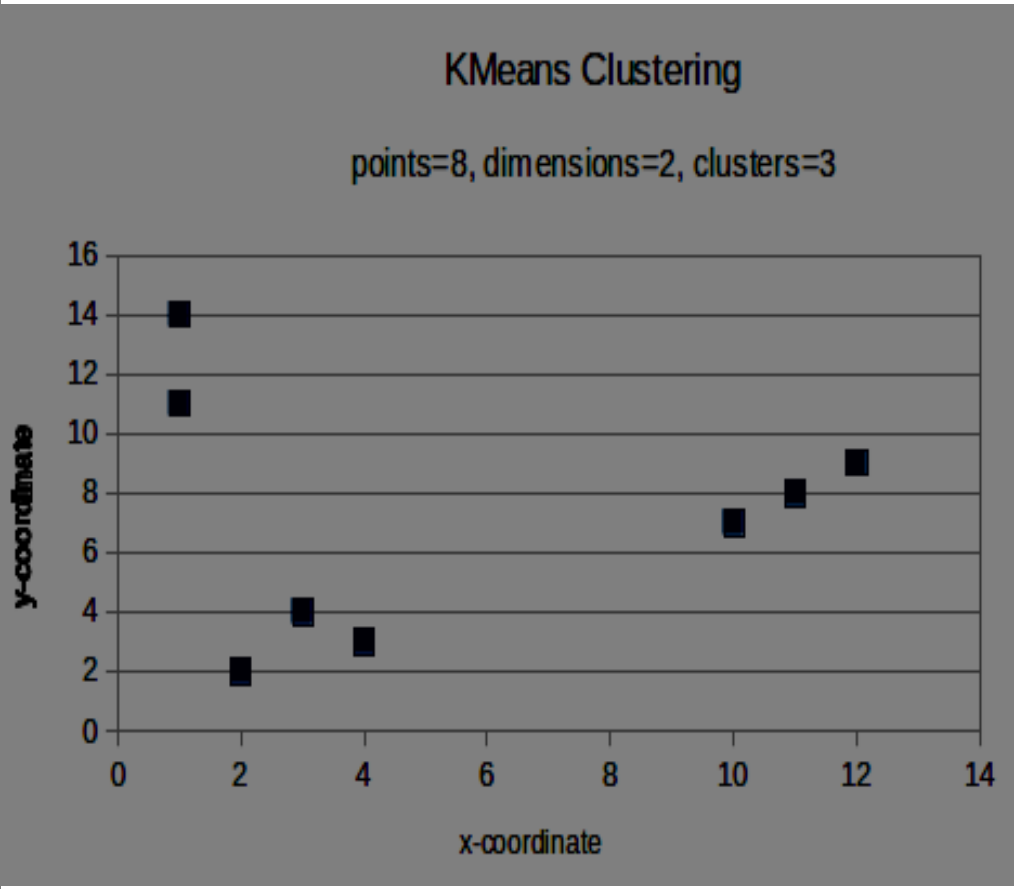
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where μ_i is the geometric centroid of data points in S_i and x_j is the vector representing jth data point.

MAPREDUCE IMPLEMENTATION

DEMONSTRATION





RESULTS AND STATISTICS

- INPUT DATA SET
  - Number of points(n) = 1 million
  - Dimensions(d) = 10
  - Number of clusters(k) = 10
  - Size of dataset = 31.2 MB
- RUN TIME ON LOCAL MACHINE
  - For #iterations = 1, runtime = 0:09:33.797533
- RUN TIME ON HADOOP
  - For #iterations = 1, runtime = 0:11:45.287343