CS543 Final Project: Dimension reduction for clustering

Ningxiao Tang
Yunshuang Tao

Project Overview

Goal: Perform dimensionality reduction of dataset and compare the cost of each dimension.

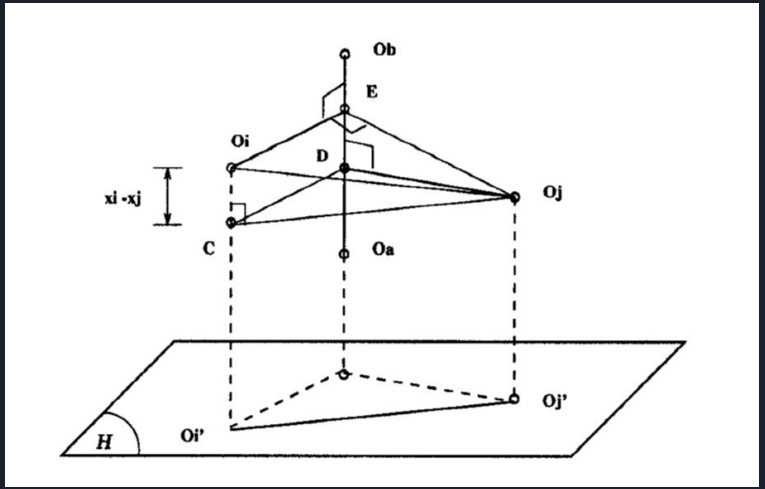
Methods:

1. Use FastMap to reduce dimension of the dataset
2. Use k-means++ to assign points into different clusters and keep the result of cost of each dimensions.
3. Compare the cost of the target dimension and the original dimension

GIthub Link: https://github.com/Ningxiao-Tang/BigDataProj

FastMap Algorithm

1. Reduce a set of data of N dimensions to k dimensions
2. Initialize column to 0 (In practice it should be -1), compute distance matrix
3. Repeat the following steps k times
4. Increment column by 1
5. Choose pivot objects Oa and Ob and record the ids of the pivot objects
6. Project objects online (Oa, Ob) and update dataset
7. Update distance matrix, decrement k by 1



Choose Distant Objects

1. Choose arbitrarily an object and let it be the second pivot object Ob
2. Let Oa be the object that is farthest apart from Ob
3. Let Ob be the object that is farthest apart from Oa
4. Report the objects Oa and Ob as the desired pair of objects

K-Means++ Algorithm

1. Randomly select first means of size N
2. Select k centers from means
3. Update cluster in while loop until the euclidean distance between new centroids and old centroids becomes zero
    a. Assign each value to its closest cluster
    b. Sorting the old centroid values
    c. Finding the new centroid by taking the average value

Cost Function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{m} a_{ij} ||x_i - \mu_j||_2^2$$

Where:

if $x_i \epsilon j$ Cluster:

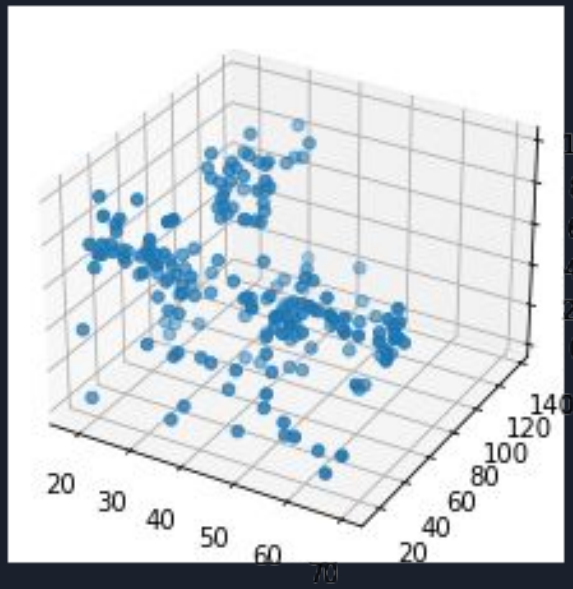
$$a_{ij} = 1$$

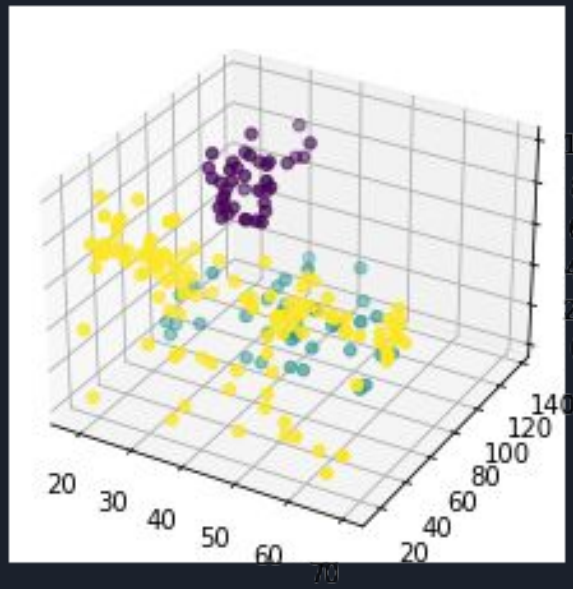
else:

$$a_{ij} = 0$$

Cost Function

Sample of Clustering





Dataset and Experiment

Original dataset: http://archive.ics.uci.edu/ml/machine-learning-databases/00401/

Choose N points and n dimensions to create a dataset of N “points” with original dimension n.

Reduce points dimension using FastMap

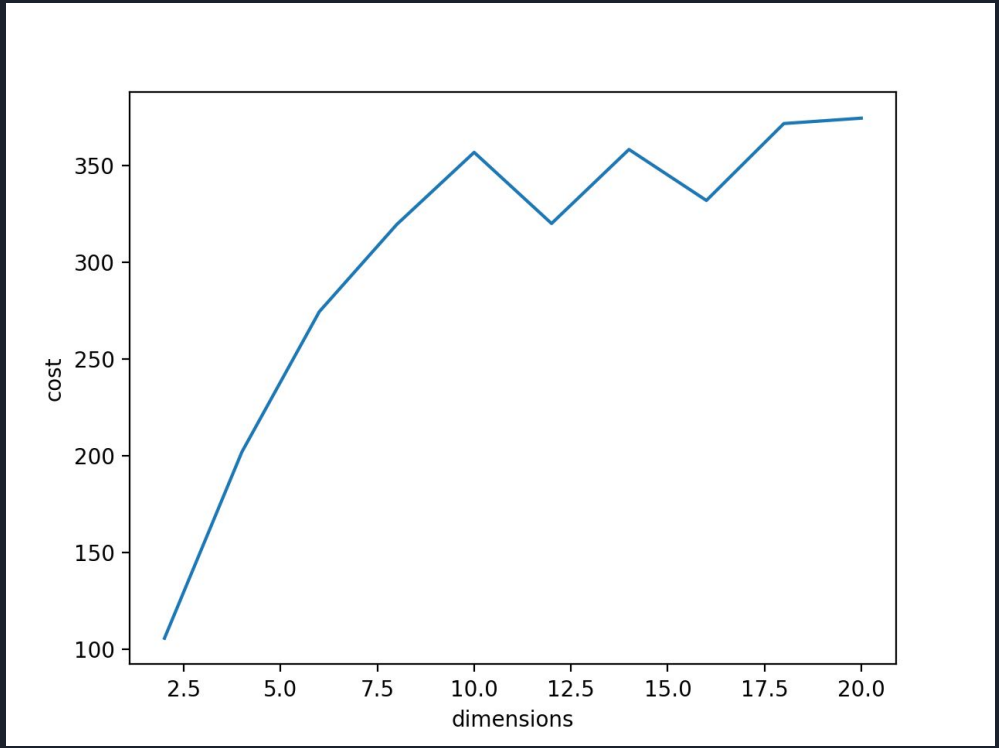
Assign points to different clusters using k-means++ clustering

Compute cost of clustered points

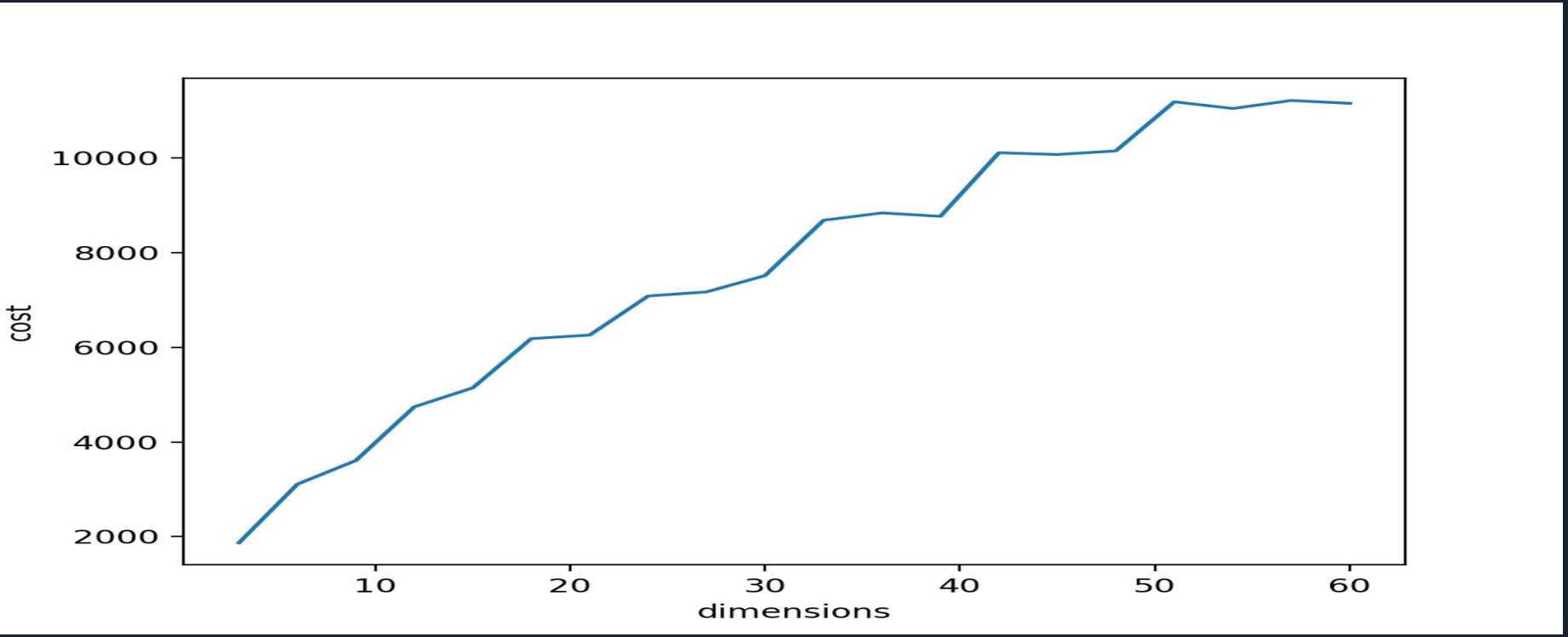
Experiment Result

- Experiment 1:
  - Number of Points = 100
  - Original Dimension = 20
  - Number of Clusters = 6
- Experiment 2:
  - Number of Points = 800
  - Original Dimension = 60
  - Number of Clusters = 16
- Experiment 3:
  - Data: image with shape (3376, 6000, 3)
  - Number of clusters: [2,3,4,5,6,7,8,9]

Experiment 1 Result



Experiment 2 Result



Experiment Results

To keep the cost of k-means++ clustering meanwhile reduce the workload, dimensionality reduction can be used.

There is a bound of what dimension can be reduced to.

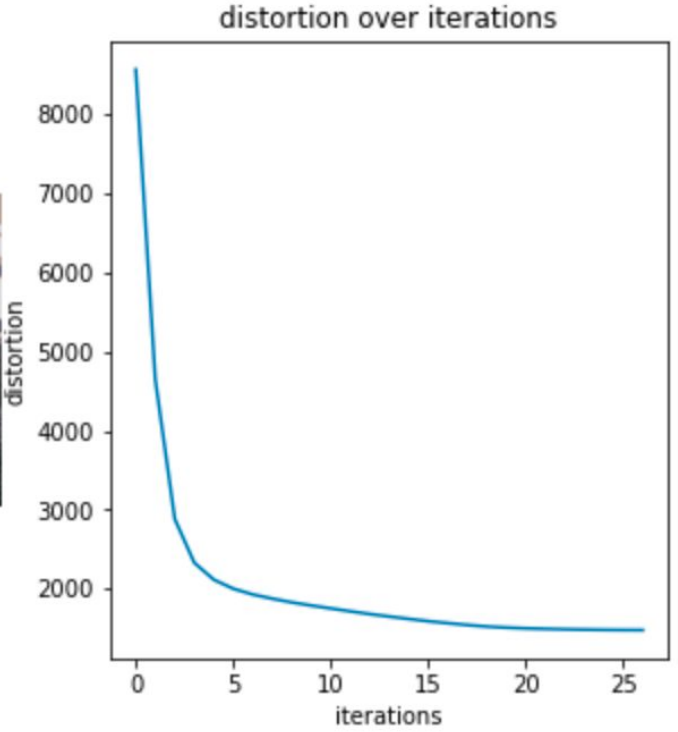
Basically dimension can be reduced to 90 percent.

If we keep reduce dimension to a large scale, cost may not be kept as the original dimension.

Experiment 3 Result
When K = 5

original image

compressed image



Elbow plot for k-means ++ algorithm