CS589 Homework #6

Trung Dang

October 25, 2023

Task 1

PCA Description

- First I mean-centered the given dataset
- Then I calculated the Covariance matrix of the dataset, denoted as X, using the formula:

$$Cov = \frac{1}{P} X \cdot X^T + \lambda I_{N \times N}$$

  while the original formula only includes $\frac{1}{P} X X^T$, as mentioned in the lecture, $\lambda I_{N \times N}$ provides more numerical stability to the equation. Additionally, λ is also chosen to be very small (10^{-7}).

- Then I calculated the eigenvalues and eigenvectors of the correlation matrix:

$$Cov = V D V^T$$

- and the eigenvectors V allows us to recover precisely the orthonormal basis we are looking for.
- Then we can calculate the encoded data as:

$$W = C^T \cdot X$$

  where C is a matrix of size $N \times K$, with $K \leq N$ be the number of basis that we are considering.
  Consequently, W is of size $K \times P$ where P is the number of data points.

Original Data

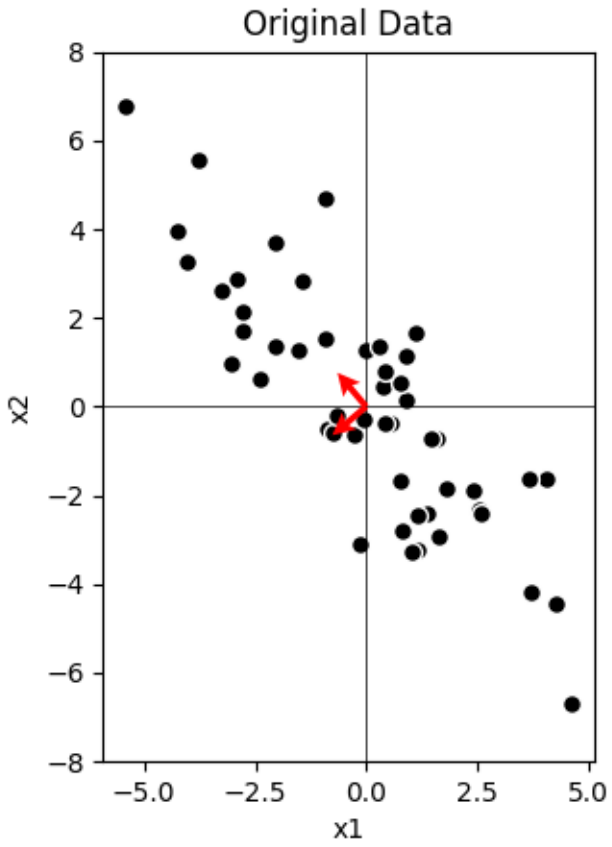

Figure 1: Original Data

Encoded Data

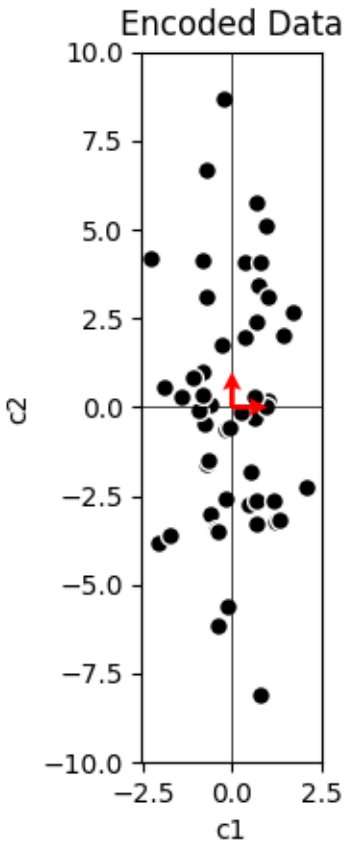

Figure 2: Encoded Data

Task 2

Centroids Initializaton

- I chose K centroids randomly in the range of the data
- The reason this method is preferable to, for instance, letting all centroids equal to the origin, is that when we calculate the distance, as the points coincide, numpy argmin will always return the smaller value (that is, if the distance to cluster 1 and cluster 2 is equal, then numpy will classify it into cluster 1). This results in a lot of empty clusters.
- Similar initialization where centroids are in close proximity also exhibits the same problem of empty clusters.
- Therefore, I found it's best to generate random centroids across the range of data inputs

$K = 3$ clustering visualization

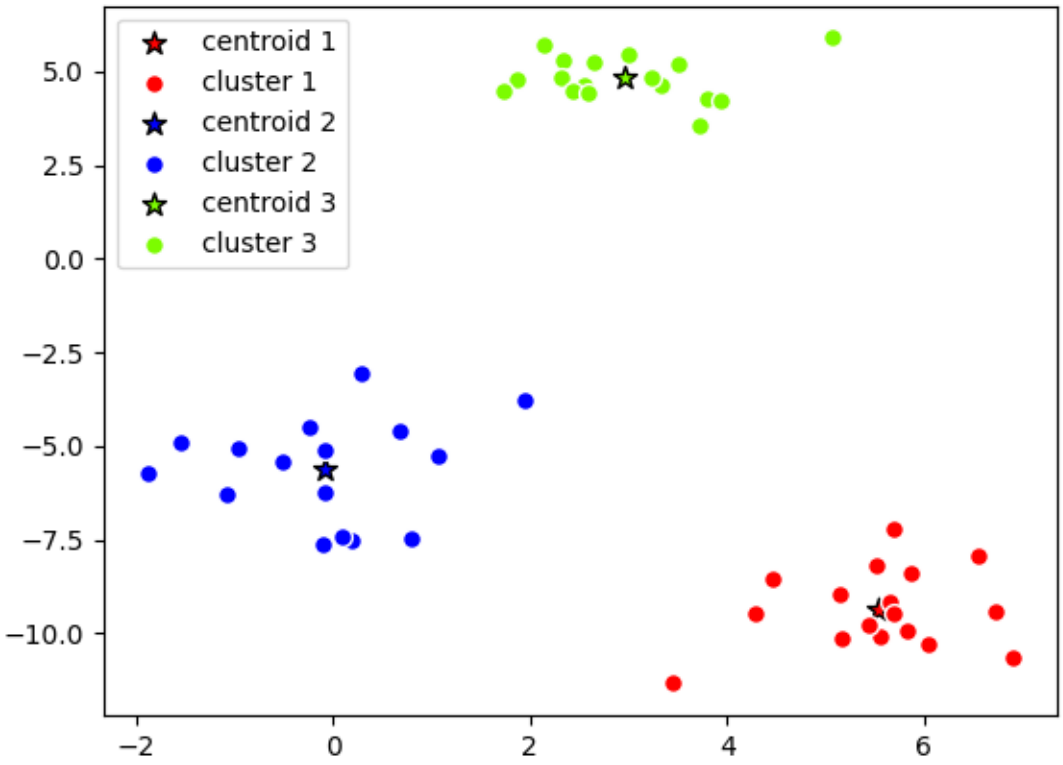

Figure 3: $K = 3$ clusters and centroids

Intra-cluster distance plot and best K

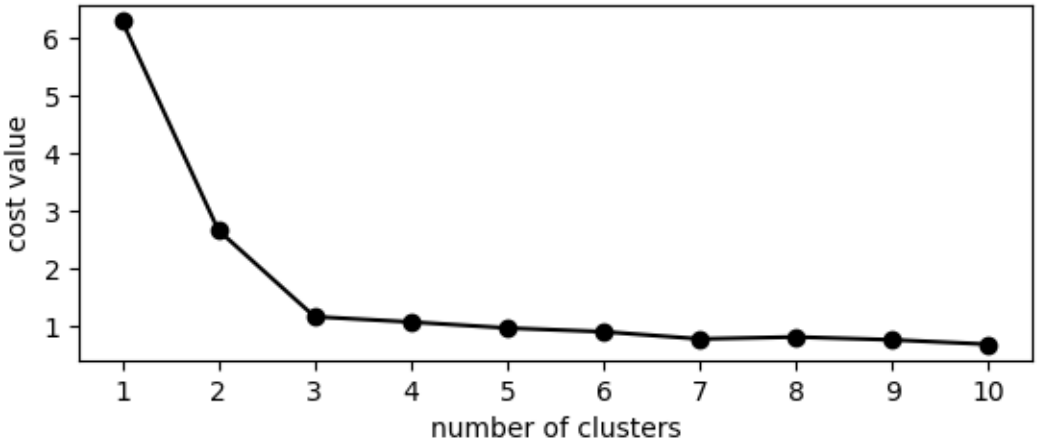

Figure 4: Task 2 scree plot with number of clusters from 0 to 10

Based on the plot, the best K for the problem is 3. The reason is that the cost value (or the intra-cluster distance) significantly decreases from $K = 2$ to $K = 3$, but does not change much from $K = 3$ to $K = 4$, nor does it decrease significantly as K gets much larger.