Homework 6

Jian Huang

2023/10/20

Task 1: Implement PCA

1. Explanation of my implementation procedure
My implementation of PCA consists of the following steps:

- Center the original data points x (x has a dimension of N*P, N: dimensional of each point/vector, P: number of data sample): $\mathbf{X} = x - \bar{x}$
- Calculate the covariance matrix of $\mathbf{X}$: $Cov(\mathbf{X}) = \frac{1}{P}\mathbf{X}\mathbf{X}^T$
- Perform eigendecomposition of the covariance matrix: $Cov(\mathbf{X}) = \frac{1}{P}\mathbf{X}\mathbf{X}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T$. The $\mathbf{V}$ is the eigenvectors and $\mathbf{D}$ is the eigenvalues.
- Perform dot product of $\mathbf{V}$ and $\mathbf{X}$ to get encoded data points w_p: $w_p = \mathbf{V} \cdot \mathbf{X}$

2. A figure that shows the mean-centered data along with its two principal components. And A figure that shows the encoded version of the data in a space where the principal components are in line with the coordinate axes. Red arrows show the principle components.

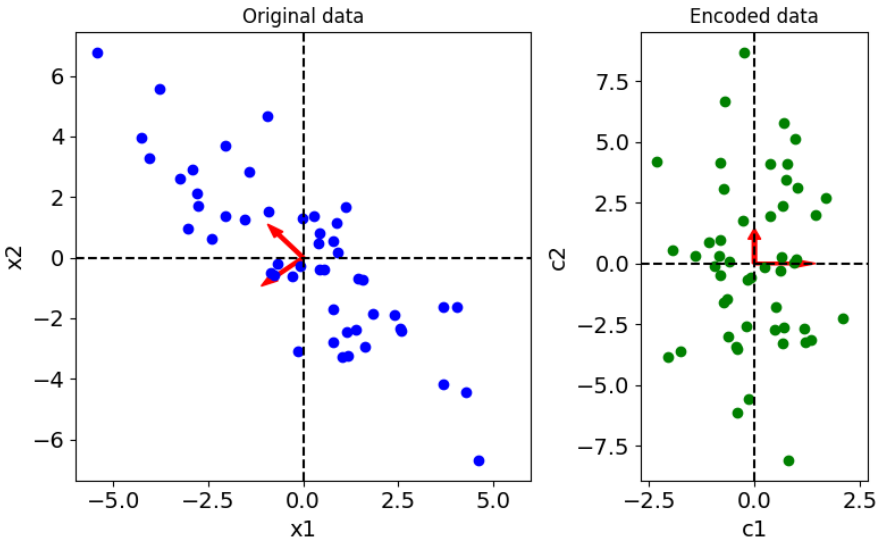

Figure 1: PCA analysis

Task 2: Implementing K-Means

1. Describe the initialization process of the centroids and reasons.

Given each data point is a two dimensional vector, the centroids are generated randomly in the range of $[x_{min}, x_{max}]$ of each dimension. As we all know, the k-means method is sensitive to the initialization of centroids. In my procedure, I added 5 iterations (repeats) and each iteration will generate a new set of randomized starting centroids. The best initialization (with the lowest cost) will be picked as the final decision.

2. A figure that visualizes your clustering results when K=3.

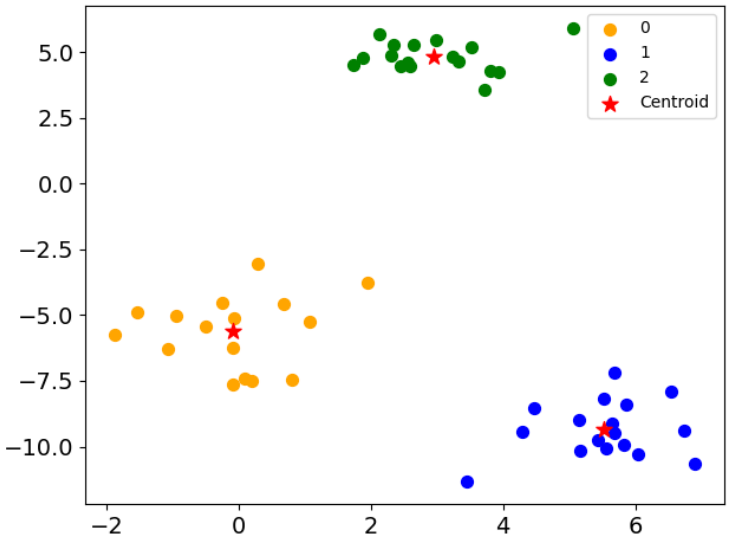

Figure 2: K-means clustering when centroid number = 3

2. A figure showing the cost with varying number of clusters from 1 to 10.

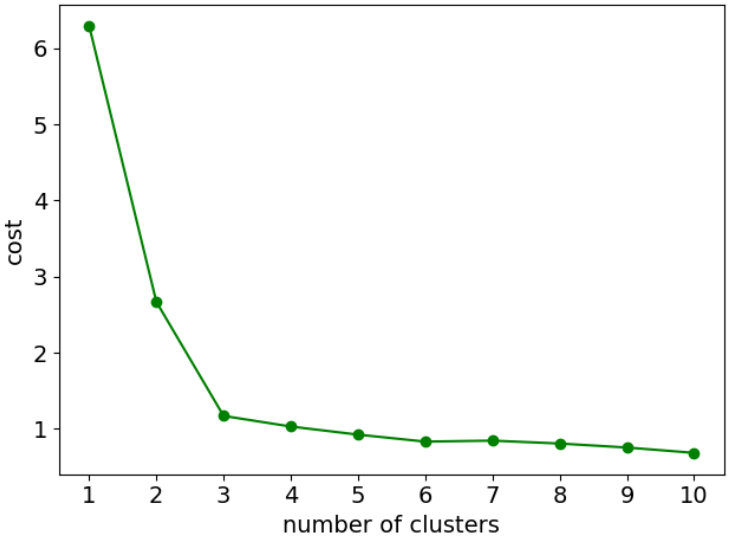

Figure 3: K-means clustering when centroid number ranging from 1 to 10

Based on Figure 3, I think K = 3 is the best option since the cost with K =3 lies in the "knee" point. With further adding more clusters, not much decreasing of the cost was gained.