COMPSCI 589 Machine Learning
Assignment 6 Report

Task 1

- Briefly describe how you implement PCA. You may include equations in your report to help with your explanation.

  I implemented PCA using the following steps:

  1. Center the data by subtracting the mean from each feature.
  2. Compute the covariance matrix of the centered data.

     Cov = 1/P * X_centered * X_centered.T
     where:
     P is the number of data points

  3. Compute the eigenvalues and eigenvectors of the covariance matrix.

     D, V = np.linalg.eigh(Cov)
     where:
     D is a vector of eigenvalues
     V is a matrix of eigenvectors

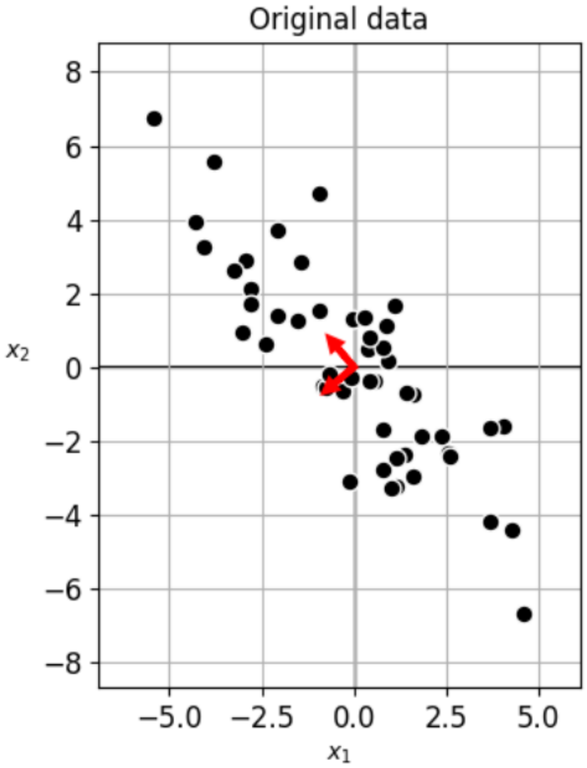
  4. Sort the eigenvectors in descending order by eigenvalue.
  5. Select the top k eigenvectors, where k is the number of principal components desired.
  6. Project the centered data onto the top k eigenvectors to obtain the principal components.

     W = np.dot(V.T, X_centered)
     where:

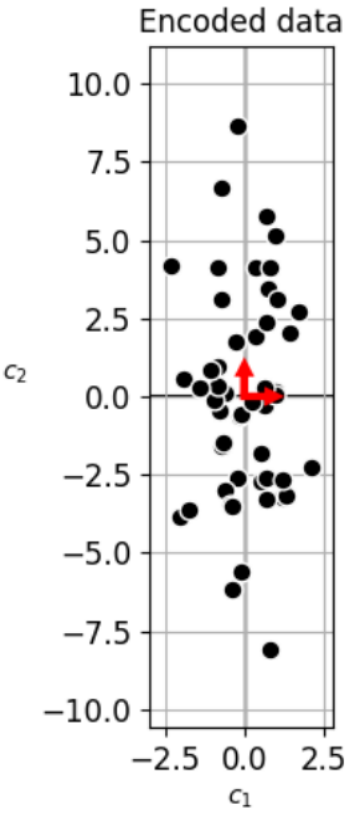
     W is the matrix of principal components

  I chose this method for implementing PCA because it is a standard and well-understood algorithm. It is also relatively efficient to implement, especially for small to medium-sized datasets.

- A figure that shows the mean-centered data along with its two principal components. In the figure, include the data points and arrows indicating the two principal components.



- A figure that shows the encoded version of the data in a space where the principal components are in line with the coordinate axes. In the figure, include the data points and arrows indicating the two principal components.

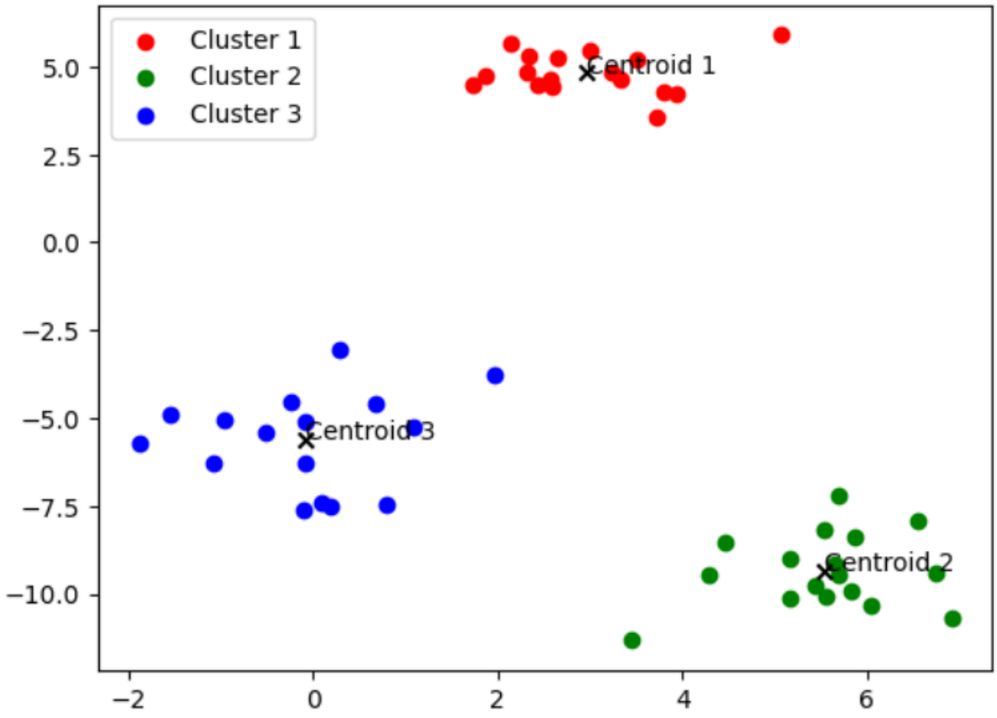


Task 2

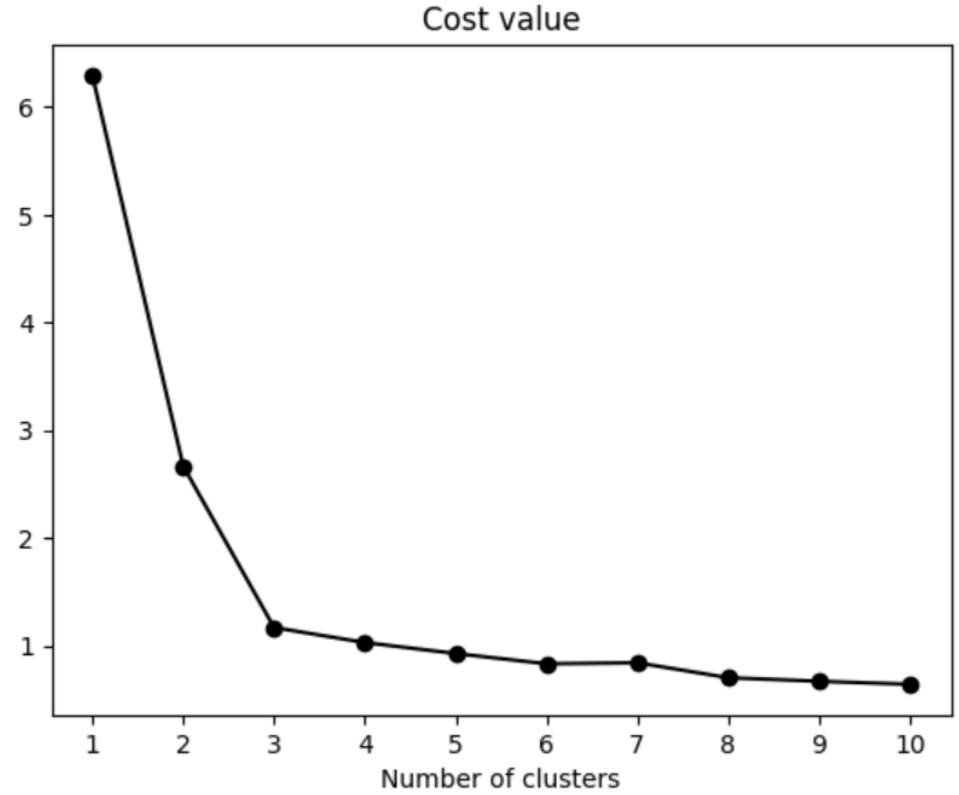
- Describe how you initialized the centroids and why you choose this method.

  I initialized the centroids as random points. This is a common method for initializing centroids, and it is relatively simple to implement. I chose this method since it's a good way to ensure that the centroids are not all initialized close together, which could lead to poor convergence of the K-means algorithm.

- A figure that visualizes your clustering results when K=3. To be more specific, you should color each cluster and put labels to indicate the clusters **in the figure** (not in the report). You should also highlight the centroid of each cluster and put labels to indicate the centroids as well.



- Scree plot by varying the number of clusters from 1 to 10. Based on the plot, answer what is the best K for this problem and why.



Best K for this problem would be K=3 since is it the "elbow point" in the curve. This is where the rate of decrease starts to slow down significantly, creating an "elbow" shape in the plot. Adding more clusters after this does not provide a substantial gain in reducing the within-cluster variance.