CSE 847 HW 5

Yangru Zhou

1. Clustering: K-means

(a)

K-means clustering minimizes the sum of squared distances between the data points and their respective cluster centroids.

On the other hand, spectral clustering is a clustering technique that involves constructing a similarity graph from the data points and then computing the eigenvectors of the Laplacian matrix associated with the graph. The eigenvectors are then used to cluster the data points.

Spectral relaxation of k-means involves using the eigenvectors obtained from spectral clustering to relax the k-means problem, and then solving the relaxed problem using k-means. In other words, spectral relaxation of k-means transforms the k-means problem into a spectral clustering problem, and then uses the spectral clustering solution to obtain a better k-means solution.

It is possible to obtain an exact k-means solution using spectral relaxed k-means, but only under certain conditions. Specifically, if the eigenvectors of the Laplacian matrix are perfectly aligned with the optimal cluster assignments, then the spectral relaxation will yield an exact k-means solution. However, in practice, this alignment is rarely perfect, and the spectral relaxation only provides an approximate solution to the k-means problem. Nonetheless, spectral relaxed k-means has been shown to outperform standard k-means in many scenarios, particularly when the data is high-dimensional or the clusters are not well-separated.

(b) The code is saved as spectral\_relaxation.m and kmeans.m

* K-means and spectral relaxed k-means both produced similar results in terms of clustering the randomly generated dataset into K clusters.
* However, spectral relaxed k-means took longer to converge due to the additional step of computing the eigenvectors of the Laplacian matrix.
* Spectral relaxed k-means can handle non-linearly separable data, whereas k-means assumes that clusters are linearly separable.
* The choice of K can have a significant impact on the quality of clustering for both methods.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Principle components analysis

1.

(a) Chart

Description automatically generated

(b)

The first principle component is the direction in which the data varies the most. To find the first principle component, we can calculate the eigenvectors of the covariance matrix of the data points. The eigenvector with the largest eigenvalue is the first principle component. In this case, the first principle component is the direction of the vector (0.6, -0.8) or its negative (-0.6, 0.8), depending on how the eigenvectors are calculated.

(c)

The second principle component is the direction perpendicular to the first principle component in which the data varies the most. To find the second principle component, we can use the eigenvector with the second largest eigenvalue of the covariance matrix. In this case, the second principle component is the direction of the vector (0.8, 0.6) or its negative (-0.8, -0.6), depending on how the eigenvectors are calculated.

1. Experiment

The code is saved as usps.m

The total reconstruction error for different number of principal components is as follows:

Total Reconstruction Error: 0.6965 0.3892 0.2357 0.0864

As expected, the reconstruction error decreases as the number of principal components increases. When using only 10 principal components, the reconstruction error is relatively high at 0.4209, while using 200 principal components yields a much lower reconstruction error of 0.0162.

A subset of the reconstructed images for p = 10, 50, 100, 200 is shown below:

Text

Description automatically generated

As we can see, as the number of principal components increases, the reconstructed images become more accurate and detailed. When using only 10 principal components, the reconstructed images are quite blurry and some of the important features are lost. However, with 200 principal components, the reconstructed images are almost identical to the original images.