CSE 847 (Spring 2023): Machine Learning— Homework 5

Instructor: Jiayu Zhou

Balija Santoshkumar
balijasa@msu.edu
https://github.com/santoshbalija/CSE847_HW5

1 Clustering: K-means

1. Elaborate the relationship between k-means and spectral relaxation of k-means. Is it possible that we obtain exact k-means solution using spectral relaxed k-means?

2. Implementation of k-means. Submit all the source code to D2L along with a short report on your observation.

   - Implement the k-means in MATLAB using the alternating procedure introduced in the class (you will not get the credit if you use the build-in kmeans function in MATLAB).
   - Implement the spectral relaxation of k-means. Create a random dataset and compare the k-means and spectral relaxed k-means.

Sol:
We assume that we have n data points $\{x_i\}_{i=1}^n \in \mathbb{R}^m$, which we organize as columns in a matrix

$$X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{m \times n}$$

The objective of K-means is reduce sum squared error(SSE)

$$q_j = \sum_{v \in \pi_j} \|x_v - c_j\|^2 \tag{1}$$

where c_j is the cluster center of the corresponding cluster Let e be the vector of all ones with appropriate length. It is easy to see that $c_j = X_j e/n_j$, where X_j is the data matrix of the j-th cluster.

SSE can transformed into

$$q_j = \sum_{j=1}^{k} \left(\operatorname{trace}\left(X_j^T X_j\right) - \frac{e^T}{\sqrt{n_j}} X_j^T X_j \frac{e}{\sqrt{n_j}} \right) \tag{2}$$

Define the n-by- k orthogonal matrix Y as follows

$$Y = \begin{pmatrix} e/\sqrt{n_1} & & & \\ & e/\sqrt{n_2} & & \\ & & \vdots & \\ & & & e/\sqrt{n_k} \end{pmatrix}$$

Then

$$Q(\Pi) = \operatorname{trace}\left(X^T X\right) - \operatorname{trace}\left(Y^T X^T X Y\right).$$

The k-means objective, minimization of $Q(\Pi)$, is equivalent to the maximization of trace $\left(Y^T X^T X Y\right)$ with Y .

In spectral relaxation of k-means is instead of using this specific expression for Y , it is possible to use any arbitrary orthogonal matrix for Y . This leads to the relaxed maximization problem

$$\max_{Y^TY=I_k} \text{trace}\left(Y^TX^TXY\right) \tag{3}$$

The first k vectors in the left singular matrix of X can produce the Y that maximizes this expression in Eq.3 we can see that both k-means and spectral k-means are trying to minimize the same error function, but spectral k-means is first trying to project the dataset into a lower dimensional space which makes it easier to capture the complex clustering structures.

The spectral-relaxed k-means become completely equivalent to k-means when the expression for Y becomes equal to the matrix mentioned in Eq. 2

I have used fisheriris data set for K-means

Regular K-means

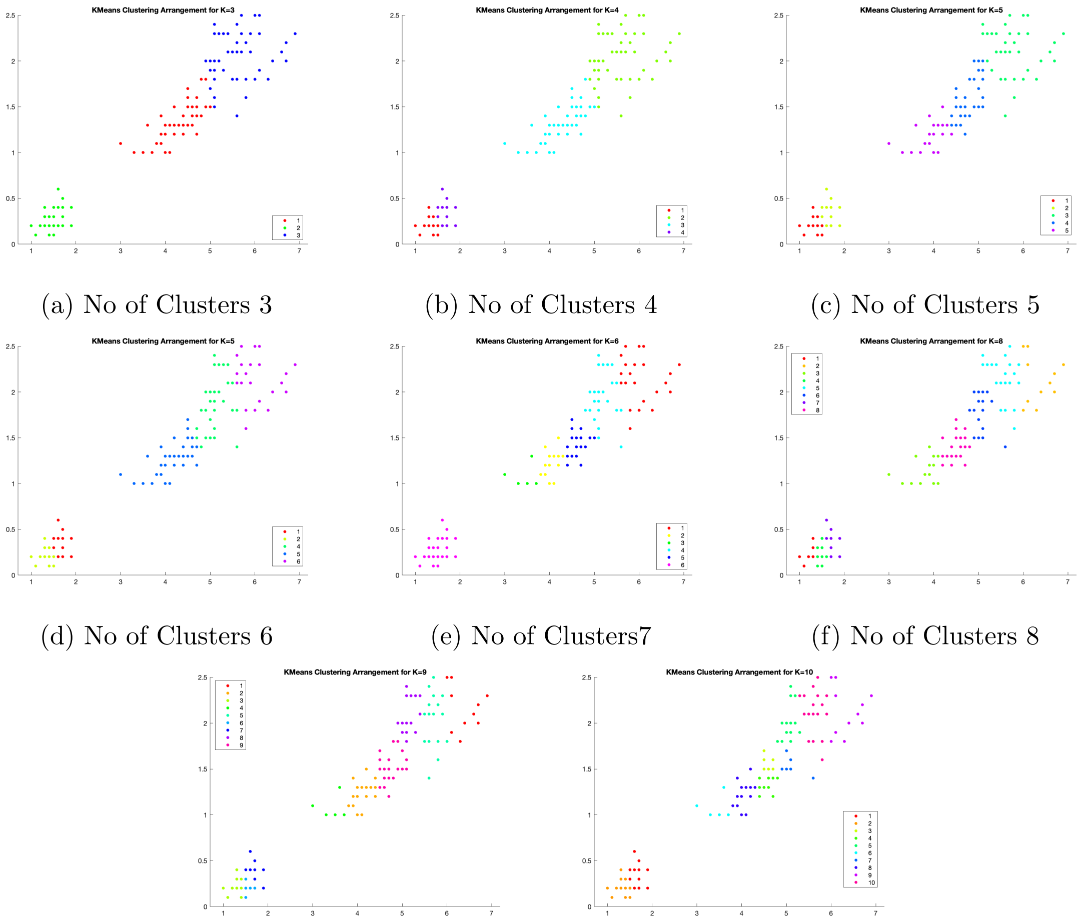
After implementation we see SSE as

```
SSE for k=3: 26.893170
SSE for k=4: 26.067859
SSE for k=5: 14.691709
SSE for k=6: 11.985356
SSE for k=7: 8.231259
SSE for k=8: 6.774484
SSE for k=9: 6.016505
SSE for k=10: 5.094189
```

spectral relaxation of *k*-means

After implementation we see Final SSE as

```
SSE for k=3: 145.287135
SSE for k=4: 107.839511
SSE for k=5: 127.624815
SSE for k=6: 82.031575
SSE for k=7: 106.387628
SSE for k=8: 70.359438
SSE for k=9: 69.009363
SSE for k=10: 43.252894
```



(a) No of Clusters 3

(b) No of Clusters 4

(c) No of Clusters 5

(d) No of Clusters 6

(e) No of Clusters7

(f) No of Clusters 8

(g) No of Clusters 9

(h) No of Clusters 10

Figure 1: Cluster assignments with K-means

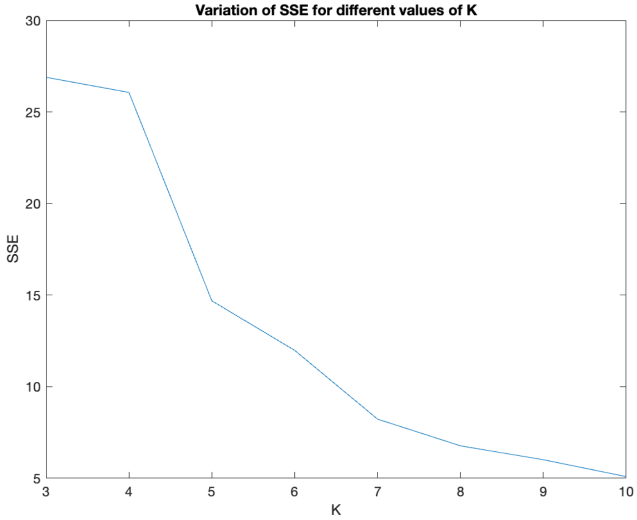
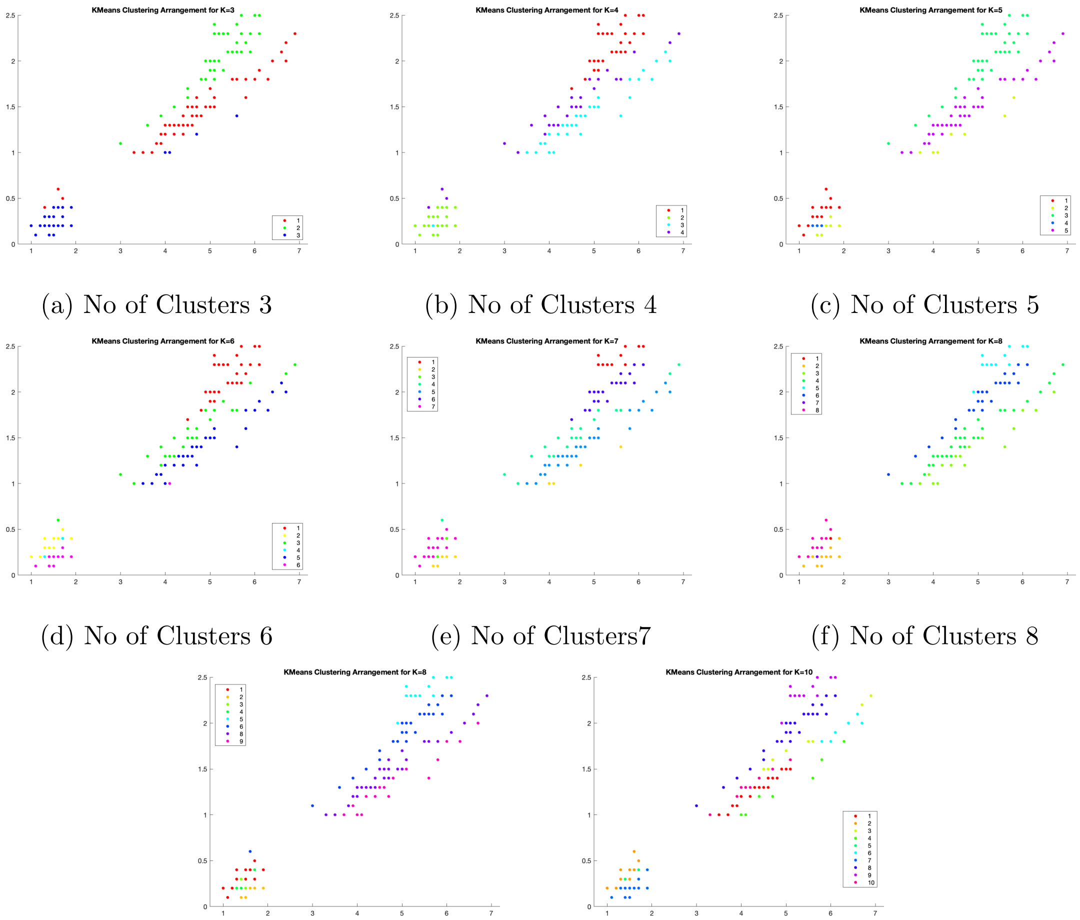


Figure 2: Convergence of SSE with Clusters variation in KMeans



(a) No of Clusters 3

(b) No of Clusters 4

(c) No of Clusters 5

(d) No of Clusters 6

(e) No of Clusters7

(f) No of Clusters 8

(g) No of Clusters 9

(h) No of Clusters 10

Figure 3: Cluster assignments with Spectral relaxation K-means

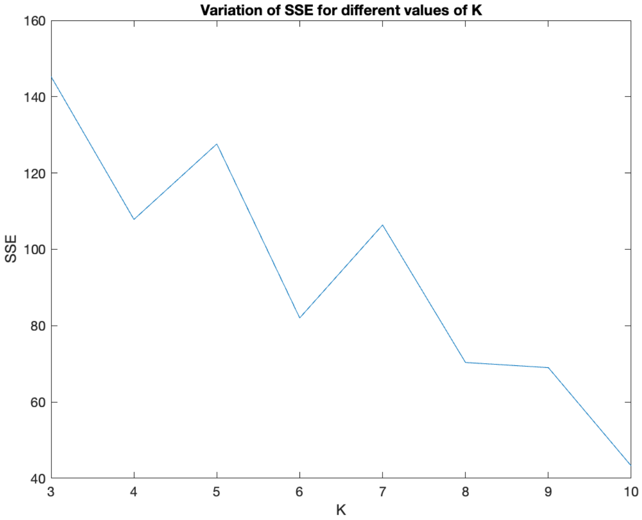


Figure 4: Convergence of SSE with Clusters variation in Spectral relaxation KMeans

2 Principle Component Analysis

1. Suppose we have the following data points in 2 d space $(0, 0), (-1, 2), (-3, 6), (1, -2), (3, -6)$.

   - Draw them on a 2-d plot, each data point being a dot.
   - What is the first principle component? Given 1-2 sentences justification. You do not need to run MATLAB to get the answer.
   - What is the second principle component? Given 1-2 sentences justification. You do not need to run MATLAB to get the answer.

2. Experiment: We apply data pre-processing techniques to a collection of handwritten digit images from the USPS dataset (data in MATLAB format: USPS.mat) [1]. You can load the whole dataset into MATLAB by load USPS.mat. The matrix A contains all the images of size 16 by 16. Each of the 3000 rows in A corresponds to the image of one handwritten digit (between 0 and 9). To visualize a particular image, such as the second one, first you need to convert the vector representation of the image to the matrix representation by $A2 =$ reshape $(A(2, :), 16, 16)$, and then use imshow $(A2')$ for visualization.

   Implement Principal Component Analysis (PCA) using SVD and apply to the data using $p = 10, 50, 100, 200$ principal components. Reconstruct images using the selected principal components from part 1.

   - Show the source code links for parts 1 and 2 to your github account.
   - The total reconstruction error for $p = 10, 50, 100, 200$.
   - A subset (the first two) of the reconstructed images for $p = 10, 50, 100, 200$. Note: The USPS dataset is available at http://www. csie.ntu.edu.tw/ cjlin/libsvmtools/ datasets/multiclass.html#usps. The image size is 16 by 16 , thus the data dimensionality of the original dataset is 256 . We used a subset of 3000 images in this homework.

Sol:

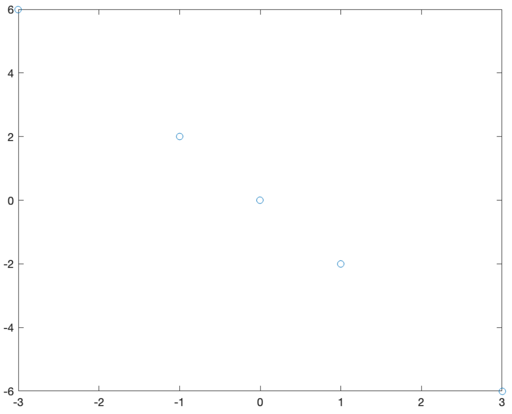


Figure 5: Data plot

By looking plot itself; we can identify first and second principal components. First component is along the data axis and second one is perpendicular to the first one.

After implementation we see reconstruction error as

```
The reconstruction error for p=10 is: 672.504293
The reconstruction error for p=50 is: 581.322126
The reconstruction error for p=100 is: 557.849767
The reconstruction error for p=200 is: 546.151961
```

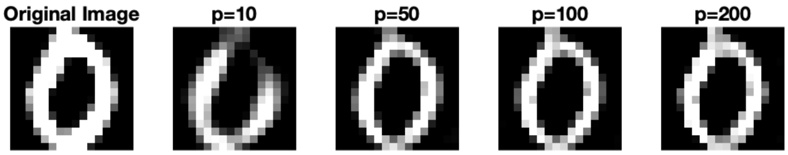


Figure 6: Visualialization of oroginal and reconstructed images PCA with different P values.