# CS182_HW4_Coding_sol

May 5, 2023

## 1 CS182_HW4_Coding.pdf is the instruction of this part.

[ ]:
```
import numpy as np
import scipy.io
import matplotlib.pyplot as plt
```

## 2 MNIST data

[ ]:
```
data_filename = 'data/images.mat'
data = scipy.io.loadmat(data_filename)
data = data['train_images']
```

### 2.1 Q1

[ ]:
```
import random
from tqdm import tqdm

Ks = [5, 10, 20]
cluster_res = {}

def KMeans(data, k, max_iter=20, stop_threshold=1e-5):
    centers = [data[:, :, np.random.randint(0, data.shape[2])] for _ in␣
 ↪range(k)]
    for _ in tqdm(range(max_iter)):
        clusters = [[] for _ in range(k)]
        for i in range(data.shape[2]):
            distances = [np.linalg.norm(data[:, :, i] - center) for center in␣
 ↪centers]
            clusters[np.argmin(distances)].append(data[:, :, i])
        pre_centers = centers.copy()
        centers = [np.mean(cluster, axis=0) for cluster in clusters]
        if np.linalg.norm(np.array(centers) - np.array(pre_centers)) <␣
 ↪stop_threshold:
            break
    return centers
```

[ ]:
```
cluster_res[5] = KMeans(data, 5)
```

```
100%|      | 20/20 [00:23<00:00,  1.19s/it]
```

[ ]:
```
cluster_res[10] = KMeans(data, 10)
```

```
100%|      | 20/20 [01:08<00:00,  3.40s/it]
```

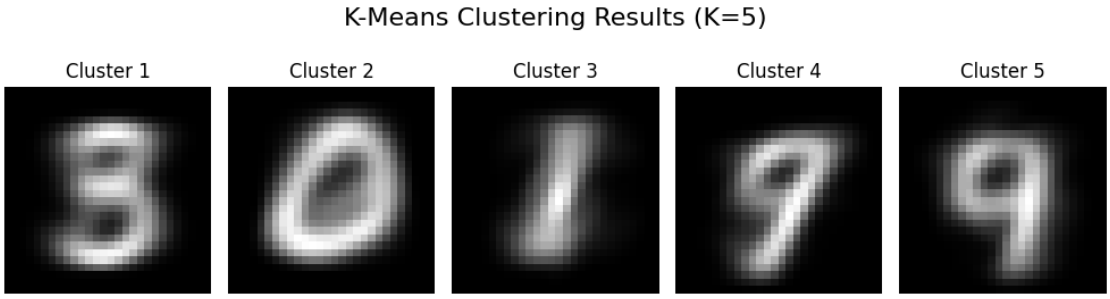
[ ]:
```
cluster_res[20] = KMeans(data, 20)
```

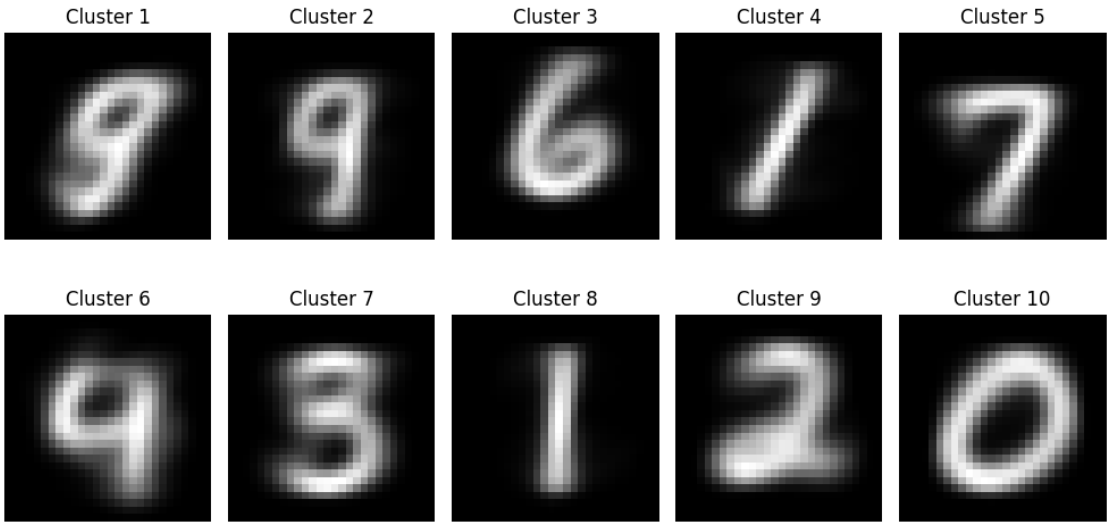
```
100%|      | 20/20 [01:20<00:00,  4.03s/it]
```

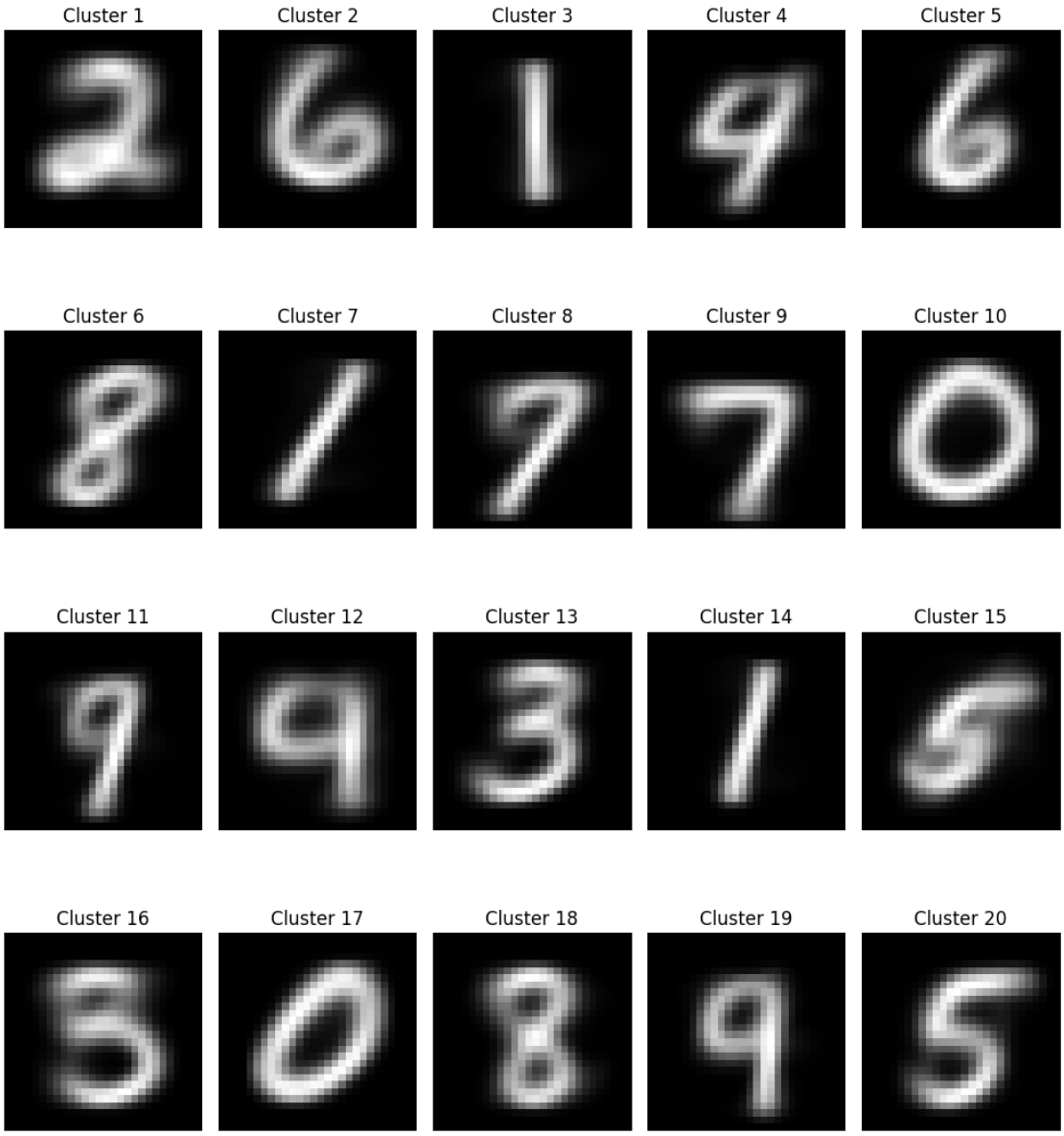
### 2.2 Q2

[ ]:
```
def plot_cluster_centers(centers, k, nrows, ncols, figsize):
    fig, axs = plt.subplots(nrows, ncols, figsize=figsize)
    axs = axs.ravel()
    for i, c in enumerate(centers):
        axs[i].imshow(c, cmap='gray')
        axs[i].set_title(f'Cluster {i+1}')
        axs[i].axis('off')
        axs[i].set_aspect('equal')
    for j in range(i+1, nrows*ncols):
        axs[j].axis('off')
    plt.suptitle(f'K-Means Clustering Results (K={k})', fontsize=16)
    plt.tight_layout()
    plt.show()

# k = 5
plot_cluster_centers(cluster_res[5], 5, 1, 5, (10, 3))
# k = 10
plot_cluster_centers(cluster_res[10], 10, 2, 5, (10, 6))
# k = 20
plot_cluster_centers(cluster_res[20], 20, 4, 5, (10, 12))
```




K-Means Clustering Results (K=10)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10




### 2.2.1 Q2 Differences between results with different numbers of cluster centers

As the result show above, with the number of cluster centers increasing, each cluster center will contains less data, which makes the result much more detailed, thus can differ data into more complex and detailed type. Meanwhile, higher cluster center number also has the problem of overfitting, so the number of cluster centers should be chosen according to the apply situation.