Machine Learning

Homework 4: Clustering And Dimensionality Reduction

Name: Parth Gandhi

Spire ID: 29693054

1. **Digit Clustering:**
   - **a.** I have used k-means clustering method. k-means clustering partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean, thus, becoming a prototype of the cluster. The equation for k-means is:

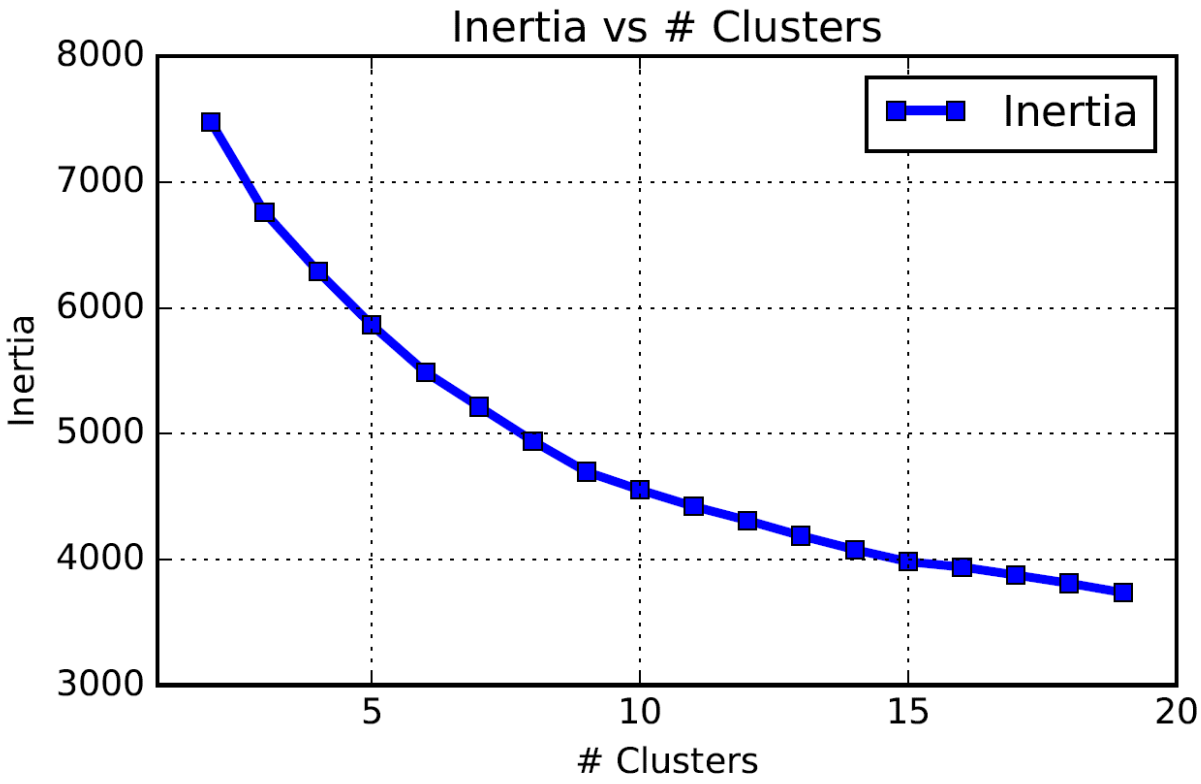
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

     K-means clustering is faster than hierarchical clustering and is also highly scalable for larger datasets. The disadvantage is that it can only work well with Euclidean distances. I chose this method because of its simplicity as well as its popularity.

   - **b.** I have used the within-cluster sum-of-squared errors(SSE) to measure the quality of clustering. It measures the 'compactness' of clusters by measuring the Euclidean distances between the data points from their assigned cluster centers. I chose this method because it is simple to implement and computationally inexpensive. The equation for SSE is:

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

c.



d. I have used the Akaike Information Criterion(AIC) to determine the optimal number of clusters in the dataset. AIC is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. It provides a relative estimate of the information lost when a given model is used to represent the process that generates the data. The equation for AIC is:

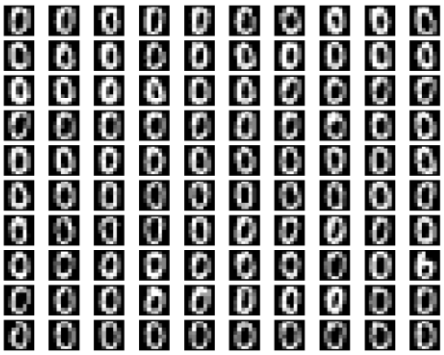
$$AIC = 2k - 2\ln(L)$$

Where, k is the number of estimated parameters in the model and L is the maximum value of the likelihood function for the model.

e.

Cluster Examples 0/11

Cluster Examples 1/11

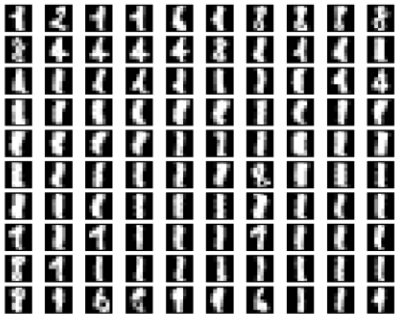


Cluster Examples 2/11

Cluster Examples 3/11

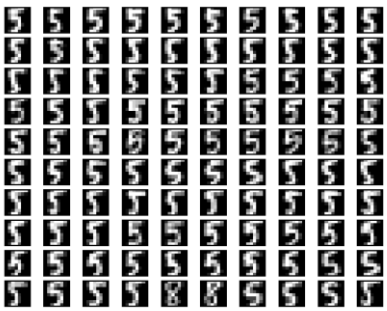
Cluster Examples 4/11

Cluster Examples 5/11



Cluster Examples 6/11

Cluster Examples 7/11

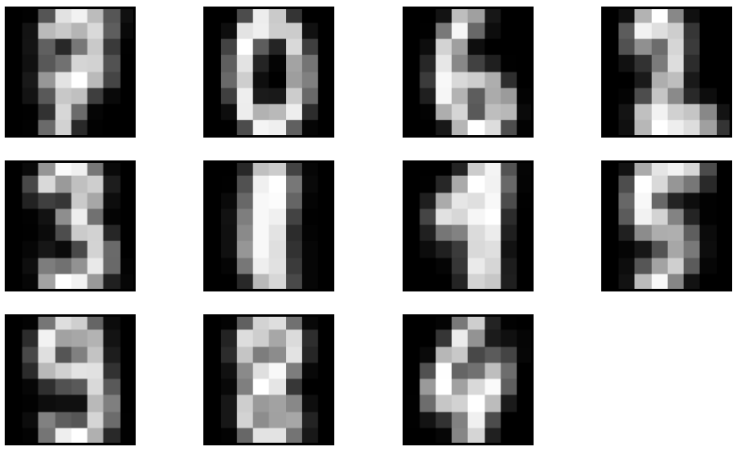


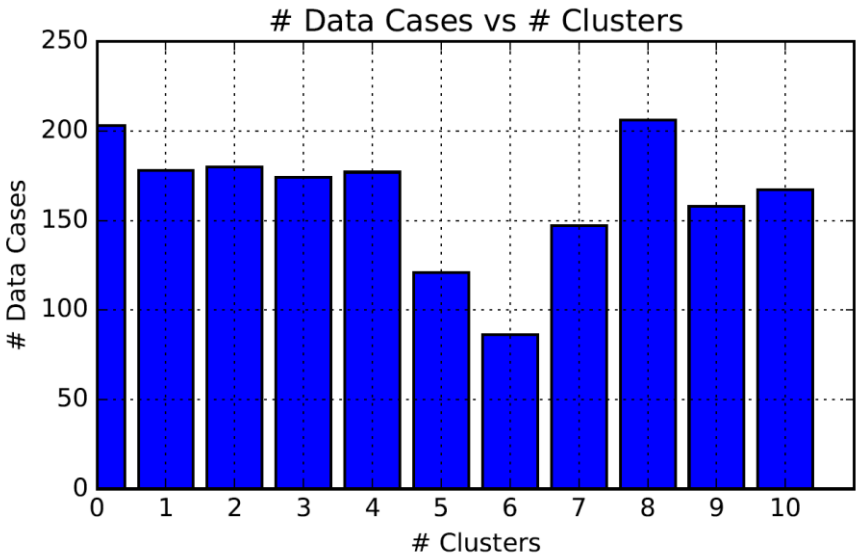
Cluster Examples 8/11

Cluster Examples 9/11

Cluster Examples 10/11

Cluster Centers



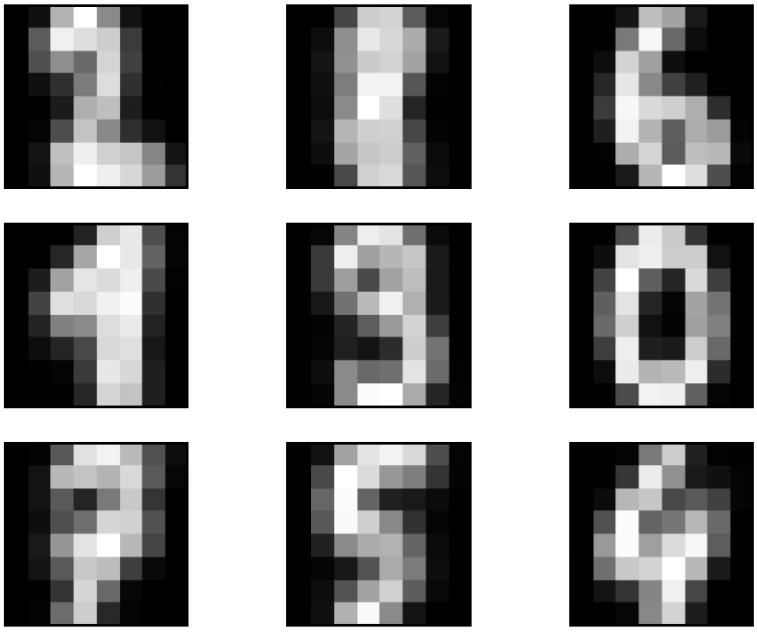


f. The number of clusters was found to be 11. It is more than the number of classes underlying the data(10). This is because the digit 1 has been classified based on two different styles.

g. Extra Credit:

I used the Bayesian Information Criterion(BIC) to determine the optimal number of clusters in the dataset. BIC is similar to AIC, the only difference being that BIC penalizes the number of parameters more strongly than AIC. The optimal number of clusters was found to be 9.

Cluster Centers BIC



The equation for BIC is:

BIC = - 2ln(L) + k.ln(n)

Where, k is the number of estimated parameters in the model, n is the number of data points in input and L is the maximum value of the likelihood function for the model.

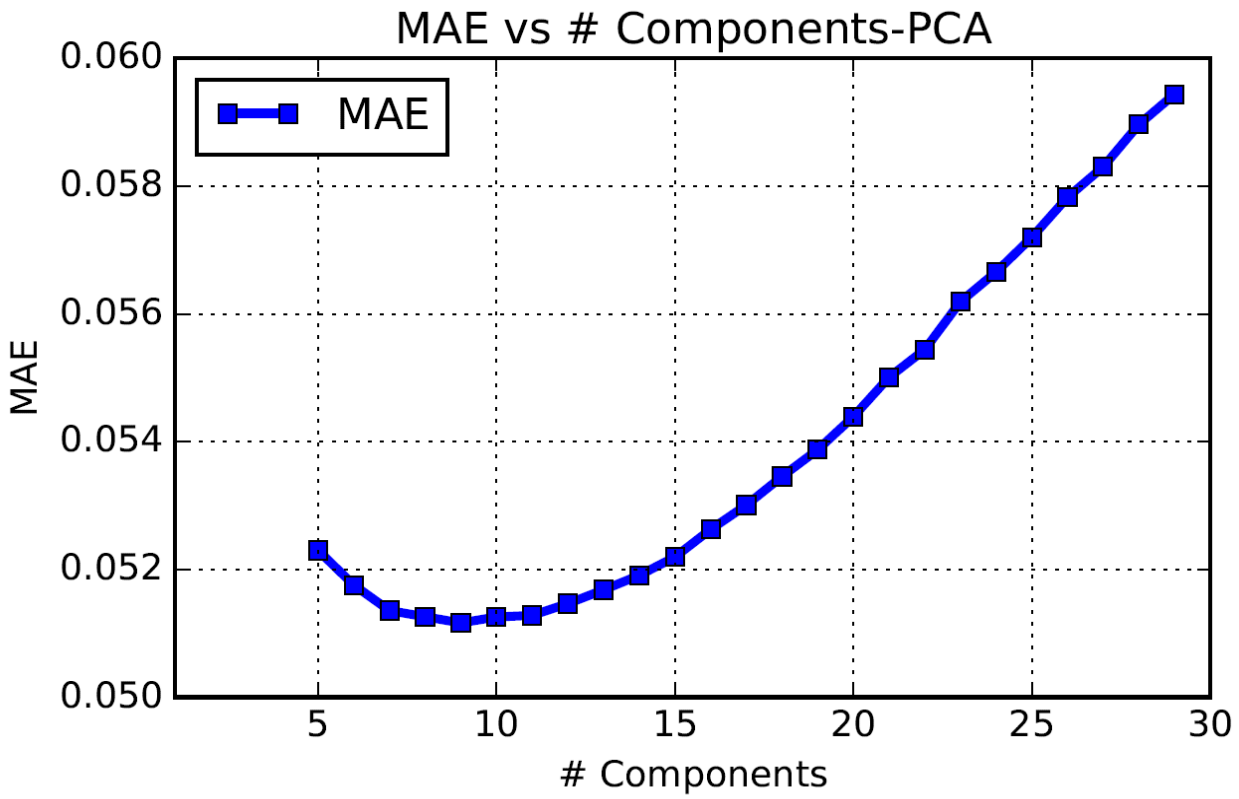
2. **Dimensionality Reduction for Image Denoising:**
   a. Principal Component Analysis(PCA) is a dimensionality reduction method that finds the dimensions with maximum variance(information) and selects the top n dimensions, thus, discarding others to achieve dimensionality reduction. It has good applications in noise removal techniques. The equation for PCA is:

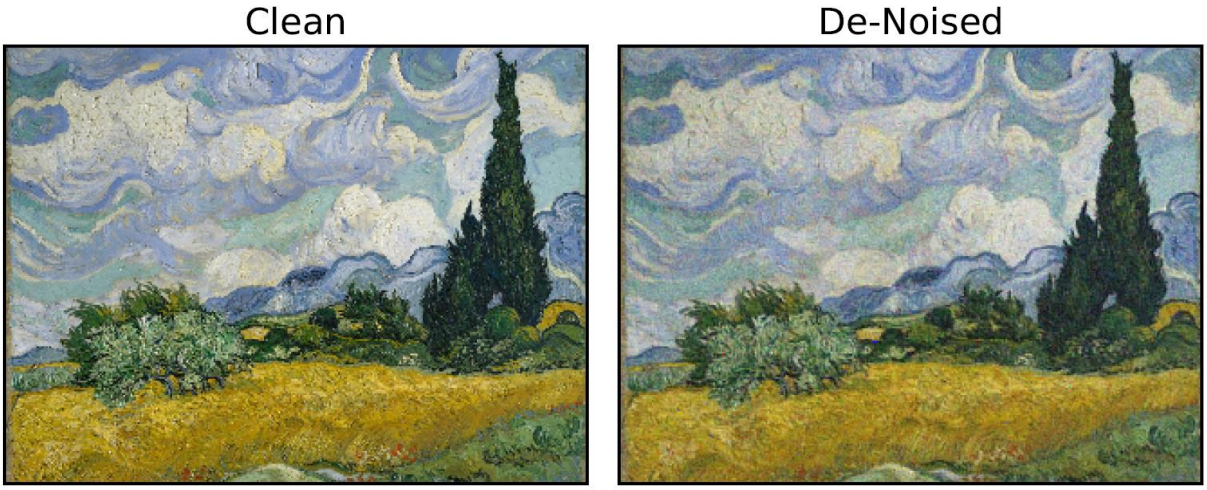
$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i\mathbf{w}-\mu)^2 \quad \text{where} \quad \mu=\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{w}$$

$$\mathbf{w}=\sum_{d=1}^{D}\omega_d\mathbf{V}_d=\mathbf{V}_1$$

Where, w is the dimension with maximum variance. The advantage of PCA is that it is a great method for denoising data like images etc. The disadvantage is that capturing invariance in data is difficult if the training dataset does not explicitly have it.

   c. The optimal number of components for PCA is 9 and the minimum value of MAE is 0.0511633287121.





d. Extra Credit:

For the extra credit, I have used the Non-Negative Matrix Factorization(NMF) dimensionality reduction method. NMF is a group of algorithms in multivariate analysis where a matrix V is factorized into two matrices W and H, with the property that all three matrices have no

negative elements. This non-negativity makes the resulting matrices easier to inspect. With NMF, the optimal number of components was found to be 11 and the minimum value of MAE was found to be 0.051381422673. The resulting plot of MAE versus the number of components as well as the plot of clean image and the optimally denoised image are shown below.