HW03

CSE575

Jean Johnson
jajoh151@asu.edu

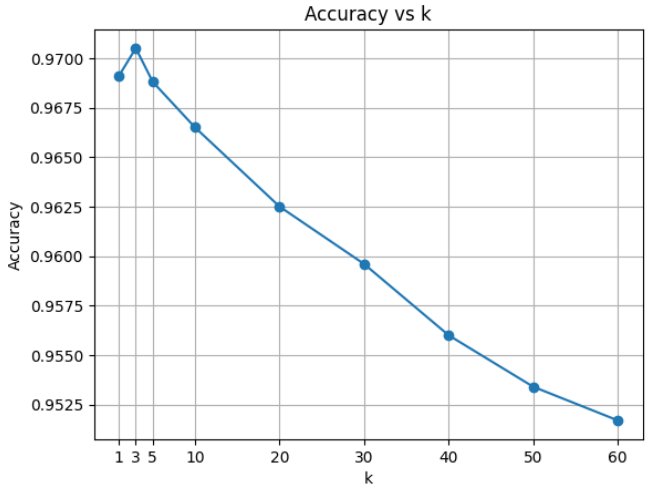
Contents

Handwritten Digits Recognition with k-NN

Prediction Accuracy

- for k = 1, accuracy is: 0.9691
- for k = 3, accuracy is: 0.9705
- for k = 5, accuracy is: 0.9688
- for k = 10, accuracy is: 0.9665
- for k = 20, accuracy is: 0.9625
- for k = 30, accuracy is: 0.9596
- for k = 40, accuracy is: 0.956
- for k = 50, accuracy is: 0.9534
- for k = 60, accuracy is: 0.9517

Plot for accuracy vs value of K

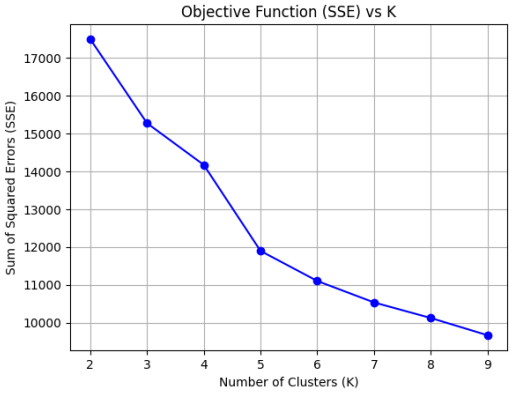


Observations

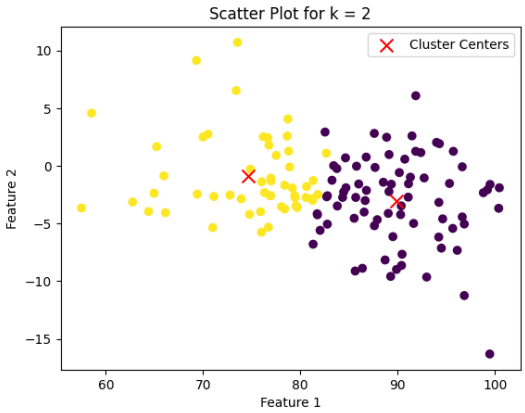
- The best prediction accuracy is at k = 3 at 97.05%
- As k gets larger, the accuracy goes down. This is because smaller k considers the data points that are the most similar and belong to the same class. As k increases, the data points that likely belongs to different classes are considered in the majority vote which can lead to a “smoothing” where the decision boundary become underfitting.

Problem 2: K-Means Clustering

Plot for objective function (SSE) vs K



Plot for cluster assignment (k=2)

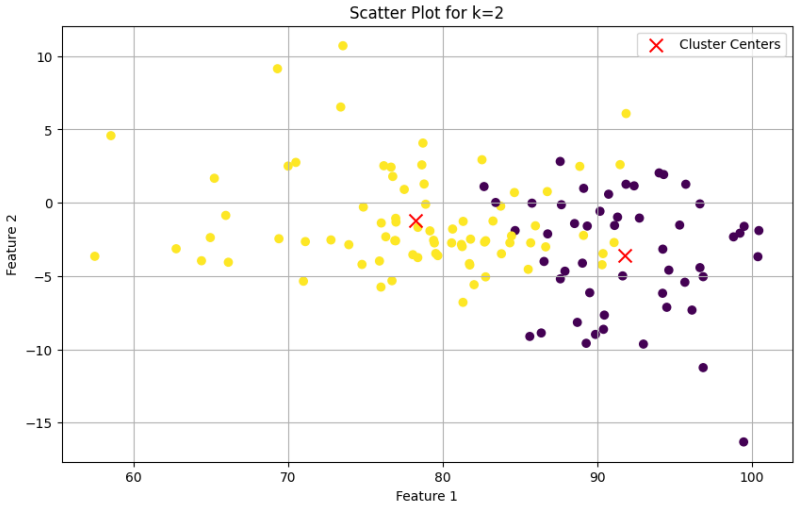


Observations

- The objective function (SSE) value decreases as value of K increases. As k decreases, distance between data points and the centroid will also decrease. When k = no. of data points, the SSE would be zero since every point is its own cluster center.
- Based on the plot, optimal number of clusters (k) is 5 since the error decrease is comparatively smaller after k = 5.
- The scatter plot shows yellow points and purple points for when we do k-means clustering for k=2. We see a few of the data points overlapping in the graph, which indicates the Feature 1 and Feature 2 are not sufficient to represent the data points in 2D.

Problem 3: Gaussian Mixture Model (GMM)

Plot for cluster assignment (k=2)



Observations

- GMM algorithm is more robust to overlapping data since it considers covariance. The decision boundary for GMM is usually curved since we do a probabilistic (soft) assignment compared to K-Means where the decision boundary is hard, straight line perpendicular to the line connecting the cluster centers.
- K-means algorithm took 6 iterations while the GMM algorithm took 38 iterations. This is because GMM is optimizing more parameters (means, covariances, mixing coefficients) compared to K-means which is optimizing just mean.