BLG 527E - Machine Learning
Ömercan Susam – 504162517

Homework 2

Q1: How to run python code: "python hw2_q1.py"

Reconstruction error and Gini Impurity values are both reduced when K number is increased in K-Means. This is because the ratio of instance number in each cluster is getting smaller if there are more centers. As a result, denser clusters are obtained with low error values.

On the other hand, when distance calculation is changed from Euclidean to Weighted, errors are again decreased since the center of the image caries more information than its edges.

Table 1 shows the average results with increased K values for Euclidean and Weighted distance calculations. Detailed results can be found in Table 2, Table 3 and Table 4 for K = 10, 20 and 30 respectively. Some clusters have no instance therefore their error is zero.

Table 1 Average Result Comparisons for Reconstruction Error and Gini Impurity results of Euclidean and Weighted Distance calculations

K	Reconstruction Error		Gini Impurity	
	Euclidean	Weighted	Euclidean	Weighted
10	289352.800	268483.500	0.284	0.223
20	134607.400	115968.350	0.140	0.163
30	69247.600	66875.833	0.100	0.151

Reconstruction error found as follows:

$$E(\{m_i\}_{i=1}^{k}|X) = \sum_t \sum_i b_i^t \|x_t - m_i\|^2$$

Gini Index found according to formula in below:

$$G = \sum_{t=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk})$$

Table 2 Reconstruction Error and Gini Impurity comparison between Euclidean and Weighted Distance calculations for K = 10

	Reconstruction Error		Gini Impurity	
	Euclidean	Weighted	Euclidean	Weighted
	0.000	53628.000	0.000	0.158
	498767.000	247503.000	0.528	0.052
	148589.000	195931.000	0.005	0.183
	227863.000	231176.000	0.089	0.094
	595581.000	450701.000	0.688	0.564
	707661.000	622983.000	0.688	0.586
	367134.000	220157.000	0.349	0.073
	0.000	228434.000	0.000	0.149
	0.000	331482.000	0.000	0.320
	347933.000	102840.000	0.493	0.053
Average:	289352.800	268483.500	0.284	0.223

Homework 2

Table 3 Reconstruction Error and Gini Impurity comparison between Euclidean and Weighted Distance calculations for K = 20

	Reconstruction Error		Gini Impurity	
	Euclidean	Weighted	Euclidean	Weighted
	211274.000	169920.000	0.013	0.613
	0.000	127578.000	0.000	0.295
	120194.000	212969.000	0.224	0.042
	0.000	143379.000	0.000	0.061
	0.000	195843.000	0.000	0.261
	0.000	30073.000	0.000	0.130
	0.000	0.000	0.000	0.000
	261098.000	141794.000	0.639	0.099
	342274.000	131904.000	0.302	0.100
	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	0.000
	680556.000	195778.000	0.716	0.067
	591259.000	157353.000	0.700	0.653
	0.000	97296.000	0.000	0.180
	147908.000	70665.000	0.000	0.010
	245298.000	243041.000	0.099	0.185
	92287.000	124863.000	0.111	0.028
	0.000	243131.000	0.000	0.486
	0.000	0.000	0.000	0.000
	0.000	33780.000	0.000	0.056
Average:	134607.400	115968.350	0.140	0.163

Table 4 Reconstruction Error and Gini Index comparison between Euclidean and Weighted Distance calculations for K = 30

	Reconstruction Error		Gini Index	
	Euclidean	Weighted	Euclidean	Weighted
	0.000	0.000	0.000	0.000
	37154.000	74595.000	0.025	0.000
	89120.000	46371.000	0.008	0.016
	0.000	76406.000	0.000	0.566
	48572.000	76441.000	0.069	0.013
	95988.000	60559.000	0.352	0.088
	157303.000	55819.000	0.216	0.068
	42026.000	32776.000	0.000	0.000
	0.000	0.000	0.000	0.000
	155233.000	87139.000	0.105	0.220
	208351.000	91121.000	0.461	0.077
	229051.000	59771.000	0.102	0.013
	79798.000	140372.000	0.277	0.588
	68820.000	54168.000	0.033	0.093
	41074.000	35419.000	0.091	0.088
	44711.000	61531.000	0.097	0.043
	46467.000	133169.000	0.072	0.059
	116462.000	52095.000	0.010	0.038
	177243.000	47385.000	0.042	0.255
	0.000	28822.000	0.000	0.160
	0.000	0.000	0.000	0.000
	34052.000	39841.000	0.066	0.000
	30033.000	168078.000	0.046	0.031
	0.000	94053.000	0.000	0.411
	95607.000	51508.000	0.109	0.279
	114769.000	75715.000	0.576	0.737
	51379.000	46836.000	0.183	0.024
	114215.000	169122.000	0.073	0.168
	0.000	45188.000	0.000	0.053
	0.000	101975.000	0.000	0.449
Average:	69247.600	66875.833	0.100	0.151

Homework 2

Q2-a: How to run python code: “python hw2_q2_a.py”

1-NN and 5-NN shows nearly the same performance however 1-NN looks slightly better when averages considered for class accuracy results. Also changes respect to distance calculation method is very low (in most case 1.695% for 5-NN class 2).
Table 5 shows the accuracy results for each class. Figure 1 and 2 depicts confusion matrices.

Table 5 Accuracy for each class with 1-NN and 5-NN classifiers

	1-NN Accuracy		5-NN Accuracy	
Class	Euclidean	Weighted	Euclidean	Weighted
0	100.000%	100.000%	100.000%	100.000%
1	99.451%	98.901%	98.901%	99.451%
2	98.870%	99.435%	97.175%	98.870%
3	97.814%	98.907%	97.268%	97.268%
4	98.343%	98.343%	98.343%	98.343%
5	98.352%	97.802%	98.352%	97.802%
6	100.000%	99.448%	100.000%	98.895%
7	98.883%	97.765%	97.207%	97.207%
8	94.253%	94.828%	91.954%	90.805%
9	93.889%	95.000%	96.111%	95.000%
Average:	97.986%	98.043%	97.531%	97.364%

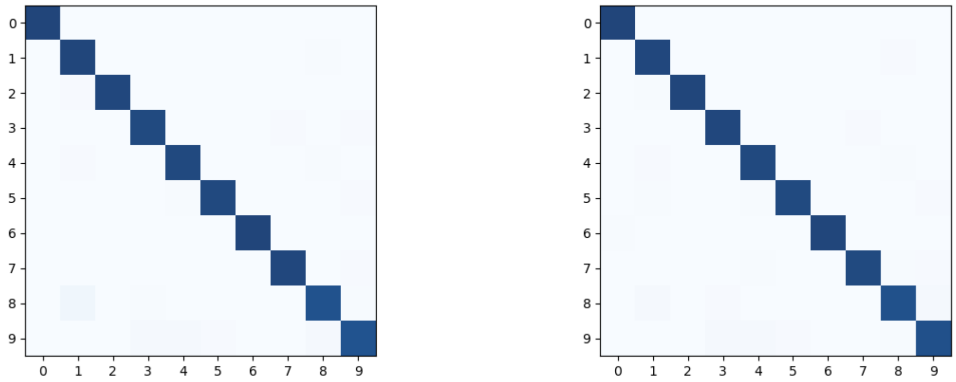

Figure 1 Confusion Matrix for 1-NN with Euclidean (left) and Weighted Distance (right).

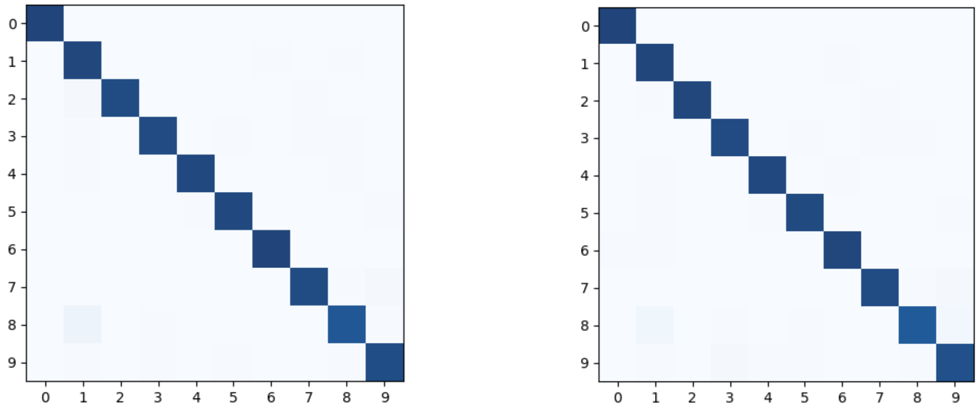

Figure 2 Confusion Matrix for 5-NN with Euclidean (left) and Weighted Distance (right).

Homework 2

Q2-b: How to run python code: “python hw2_q2_b.py”

This method is very similar with the method in the literature called Nearest Cluster Classifier (NCC). Table 6 shows the class accuracy results. Average values for distance methods does not look very differentiable. However, Class 5 classification is approximately 30% decreased when distance calculations are based on Euclidean distance.

Table 6 Accuracy for each class with k = 30

Class	Euclidean	Weighted
0	98.876%	99.438%
1	83.516%	80.220%
2	92.090%	88.136%
3	82.514%	84.699%
4	91.713%	81.768%
5	60.989%	**91.209%**
6	97.238%	95.580%
7	91.620%	86.034%
8	81.609%	79.885%
9	91.667%	94.444%
Average:	87.183%	88.141%

Table 7 shows the average class accuracy results for 1-NN, 5-NN and Centroid Based with K=30. It is easy to see that k-NN methods has better performance when compared with Centroid Based method.

Table 7 Average accuracy comparison for 1-NN, 5-NN and Centroid Based K=30

Method	Euclidean	Weighted
1-NN	97.986%	98.043%
5-NN	97.531%	97.364%
Centroid Based K=30	87.183%	88.141%