Austen Hsiao
Assignment 4 Machine Learning
985647212

Experiment 1

Of 5 runs, the lowest average *mean square error* is 643.737214697591.

For the same trial,

Mean square separation: 1172.1925342925265

Mean entropy: 0.9715052940059584

Classification accuracy for the test data: 0.7451307735114079

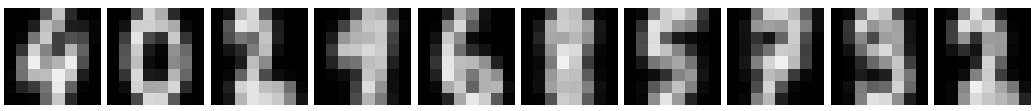
Confusion matrix for test data:

		Predicted Class									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	176	0	0	0	2	0	0	0	0	0
	1	0	58	23	1	0	0	2	0	98	0
Actual	**2**	1	1	161	0	0	0	0	2	12	0
Class	**3**	0	0	2	163	0	1	0	9	8	0
	4	0	5	0	0	162	0	0	6	8	0
	5	0	1	0	26	1	153	1	0	0	0
	6	1	0	0	0	1	0	176	0	3	0
	7	0	5	0	0	1	4	0	166	3	0
	8	1	8	1	34	0	3	2	1	124	0
	9	0	23	0	145	0	6	0	5	1	0

- On this particular trial, ‘9’ did not appear most frequently in any of the clusters, so it could not be predicted

Visualization data:

In my script, the order of the clusters’ classes as computed is: [4, 0, 2, 1, 6, 8, 5, 7, 3, 2]. In other words, the 0th cluster was found to have the highest number of 4s, cluster 1 was found to have the highest number of 0s, so on and so forth. **In the same order, the visualization data is presented below**. Note that this has no bearing on the order of the rows and columns of the confusion matrix.



Discussion:

The classification accuracy of ~75% in the test set indicates K-means clustering as a quick and effective method for elucidating hand-written digits. The visualized cluster centers look like the digits they represent with one notable inconsistency—the 4^{th} digit above looks like a '4' but was identified as a '1'. Also, the digit '9' was not identified as a cluster label for any of the clusters. The digit '9' seems to have been confused with the digit '3', based on data from the confusion matrix. When the predicted class is '3', the frequency of the actual class being '9' was less than 10% smaller than the frequency of the actual class being '3'.

Experiment 2

AverageMSE: 486.3288401524133

MSS: 1410.6401569832967

MeanEntropy: 0.31642073234244616

Accuracy on test set: 0.9248747913188647

Confusion Matrix on test set:

		Predicted Class									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	177	0	0	0	1	0	0	0	0	0
	1	0	176	0	0	0	1	0	0	0	5
Actual	**2**	0	2	172	0	0	0	0	2	1	0
Class	**3**	0	2	2	159	0	1	0	2	7	10
	4	0	5	0	0	172	0	0	1	3	0
	5	0	0	0	0	1	172	0	0	0	9
	6	0	4	0	0	1	1	173	0	2	0
	7	0	0	0	0	0	0	0	165	9	5
	8	0	26	0	0	0	1	1	5	139	2
	9	0	1	0	0	0	1	0	11	9	158

Visualization data:

The order of the clusters' classes as computed is:

[4, 1, 3, 2, 8, 2,

2, 0, 1, 1, 5, 7,

3, 0, 5, 4, 1, 8,

9, 8, 7, 8, 4, 6,

7, 5, 1, 2, 9, 9]. The visualization data is presented below.

Discussion:

The stats have improved across the board for the 30 cluster trial when compared to the 10 cluster trial. Average mean squared error is lower, indicating less distance between points and their respective centers. Mean square separation between centers has increased, a positive sign good spacing between clusters. Mean entropy has decreased, indicating less variability in the points' actual classes within each cluster. The accuracy has increased from ~75% in the 10 cluster run to about 92% in the 30 cluster run.

In the confusion matrix, it is apparent that the predicted classes matched well with the actual classes. Unlike the 10-cluster trial, the 30-cluster trial confusion matrix exhibited no competition between multiple actual classes within any given predicted class.

The visualized data, however, still has one inconsistency between the class and the image. Index 11 was computed as clustering class '7', but the image clearly shows a '9'. However, the presence of 1 inconsistency among a sample size of 30 has less impact than 1 inconsistency within a lower sample size.