Cody Herberholz
CS445
3/14/2017

K-Means Clustering

K = 10

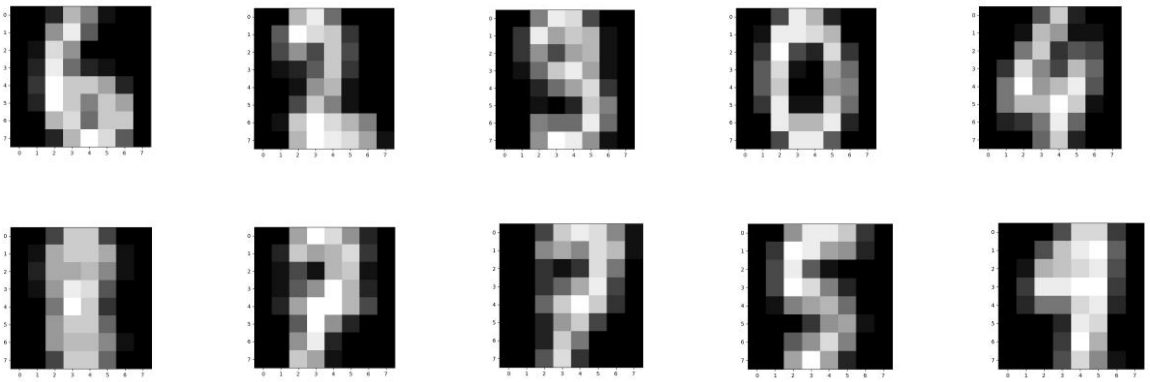
The task of the first experiment was to implement a K-means clustering algorithm using Euclidean distance with 10 clusters. Run the clustering program on the training data 5 times, keeping the 10 final cluster centers from the run with the lowest average mean-square-error.

Average mean-square-error: 6650.35

Mean-square-separation: 1321.93

Accuracy: 0.16917

Imagery



Confusion Matrix

Predicted Class

Actual Class	0	1	2	3	4	5	6	7	8	9
0	0	0	0	176	2	0	0	0	0	0
1	3	21	1	0	0	102	0	0	0	55
2	0	148	10	1	0	13	1	2	0	2
3	0	0	165	0	0	8	2	7	1	0
4	0	0	0	0	164	2	9	3	0	3
5	1	0	31	0	1	0	0	0	148	1
6	175	0	0	1	1	4	0	0	0	0
7	0	0	0	0	0	1	93	84	0	1
8	1	1	31	1	0	126	2	1	3	8
9	0	0	145	0	0	1	1	6	5	22

Discussion

When visualizing the data given from the centroids, center of the cluster, the digits displayed were recognizable despite having a poor accuracy. This might have been cause by a bug in the code since with the given accuracy the digits should have come out unrecognizable. There was a bit of confusion shown when trying to classify 8, they were always interpreted as the number 5 and vice versa. The confusion matrix above shows clustering was successful but around the wrong numbers.

K = 30

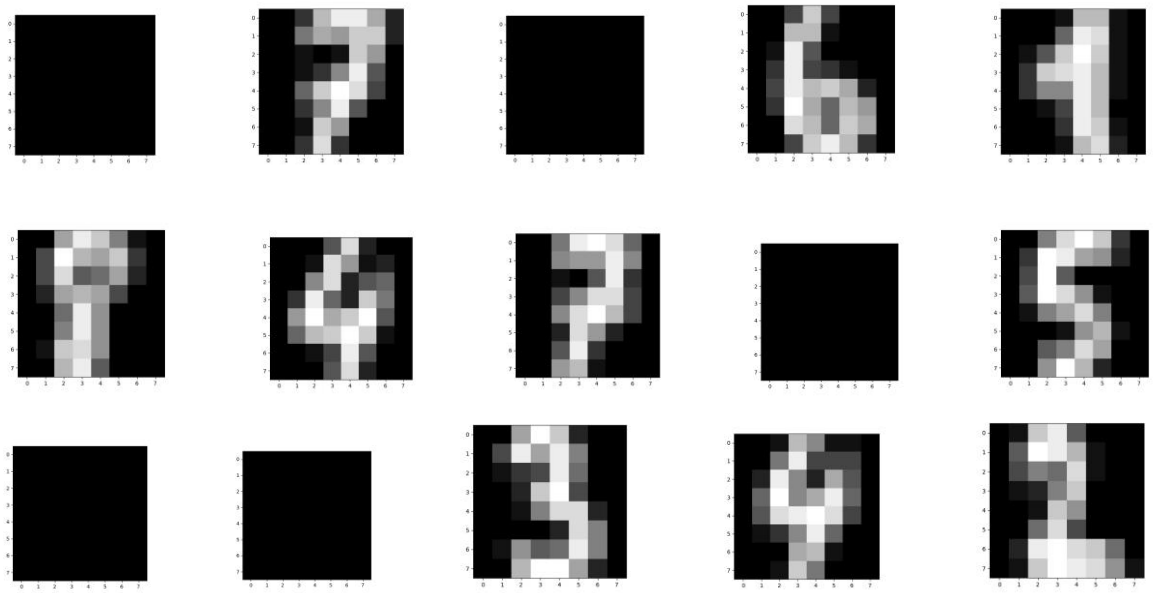
The task of the second experiment was to implement a K-means clustering algorithm using Euclidean distance with 30 clusters. Run the clustering program on the training data 5 times, keeping the 30 final cluster centers from the run with the lowest average mean-square-error.

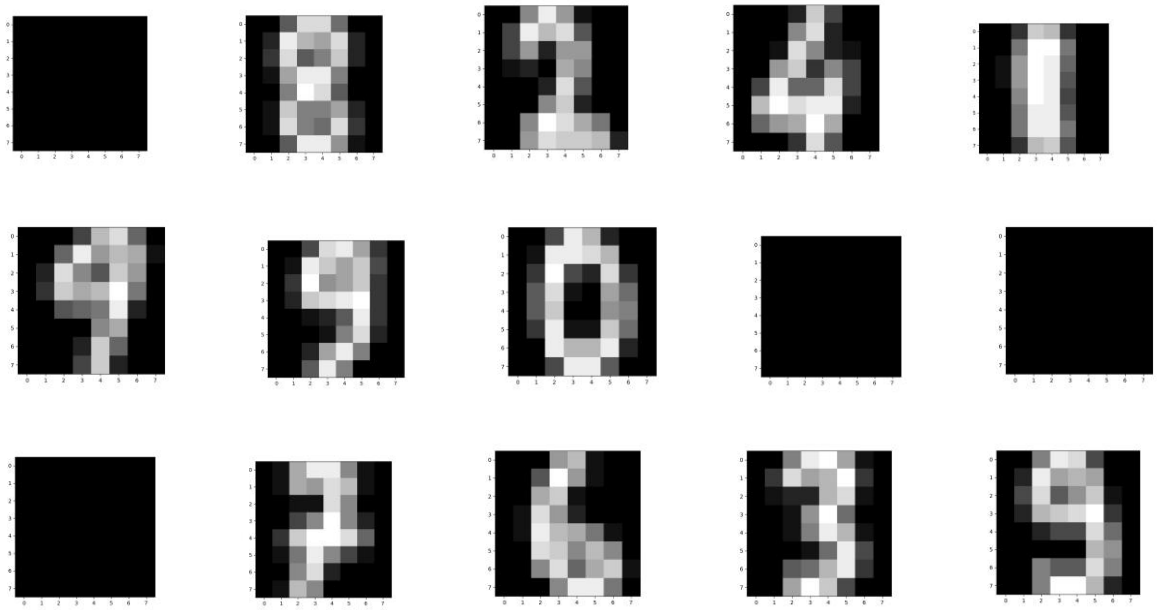
Average mean-square-error: 16275.45

Mean-square-separation: 2038.65

Accuracy: 0.04897

Imagery





Confusion Matrix

		Predicted Class																													
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**
Actual Class	**0**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	177	0	0	0	0	0	0	0
	1	0	0	0	0	53	2	0	0	0	0	0	0	3	0	17	0	0	0	0	97	4	0	0	0	0	0	0	6	0	0
	2	0	3	0	0	2	0	0	0	0	0	0	0	0	0	92	0	6	68	0	4	0	0	1	0	0	0	0	0	0	1
	3	0	2	0	0	0	4	0	1	0	2	0	0	79	0	0	0	3	2	0	0	0	0	0	0	0	0	0	0	14	76
	4	0	0	0	0	3	3	90	0	0	0	0	0	0	57	0	0	0	0	25	3	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	35	1	0	0	102	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	29	0
	6	0	0	0	82	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	1	0	0	0	0	94	0	0
	7	0	56	0	0	0	6	0	50	0	0	0	0	0	1	0	0	0	0	0	0	5	0	0	0	0	0	61	0	0	0
	8	0	1	0	0	2	27	0	0	0	0	0	0	1	1	0	0	114	0	0	20	3	4	0	0	0	0	0	1	0	0
	9	0	0	0	0	1	5	0	2	0	0	0	0	2	0	0	0	2	0	0	0	21	14	0	0	0	0	0	0	133	0

Discussion

When visualizing the data given from the centroids, center of the cluster, the digits displayed were recognizable despite having a poor accuracy. This might have been cause by a bug in the code since with the given accuracy the digits should have come out unrecognizable. When compared to experiment one the visualized digits are shown to be more clear which is definitely noticeable with the 8 above. The extra clusters allow for the data to be more spread out. If there was any conflict with the algorithm thinking a data point might belong to one of two different groups, there is now a greater chance that the clusters are farther apart and can

prevent confusion like that. One downfall of the extra clusters is that it is harder to classify the number via the usual means as can be seen from the confusion matrix above.