Homework 5: K-Means Clustering

CS 545: Machine Learning, Winter 2017

Ben Wilson

Introduction:

K-means clustering is a classifier that tries to locate and partition regions in the feature space that have are heavily clustered (i.e, have similar features). For each cluster, a centroid is found - similar to a center of mass - that minimizes the distance from each sample to the centroid. In this assignment we used the optdigits dataset to define a 64 dimensional feature space that has ten classes corresponding to the digits 0-9.

Experiment 1:

For the first experiment we are asked to use ten centroids to classify the ten digits, 0-9. While this seems like the most obvious number of centroids, it turned out to be very difficult to capture the characteristics of each digit. In most runs, the classifier couldn't find a cluster that had a majority of nines. Occasionally, it would classify a cluster as nines, but would misclassify another digit (usually threes). As a result of this, the accuracy of the classifier was typically between 70 - 74 percent.

The following is a summary of the results along with the gray-scale representation of each centroid (on the following page). In most cases the visualized centers look like their associated digits, but you can see that it didn't have a class for nines and would typically classify a nine as a three or a seven. It can also be seen from the visualized digits that it created two clusters around six, meaning it could classify only 9 of the 10 classes.

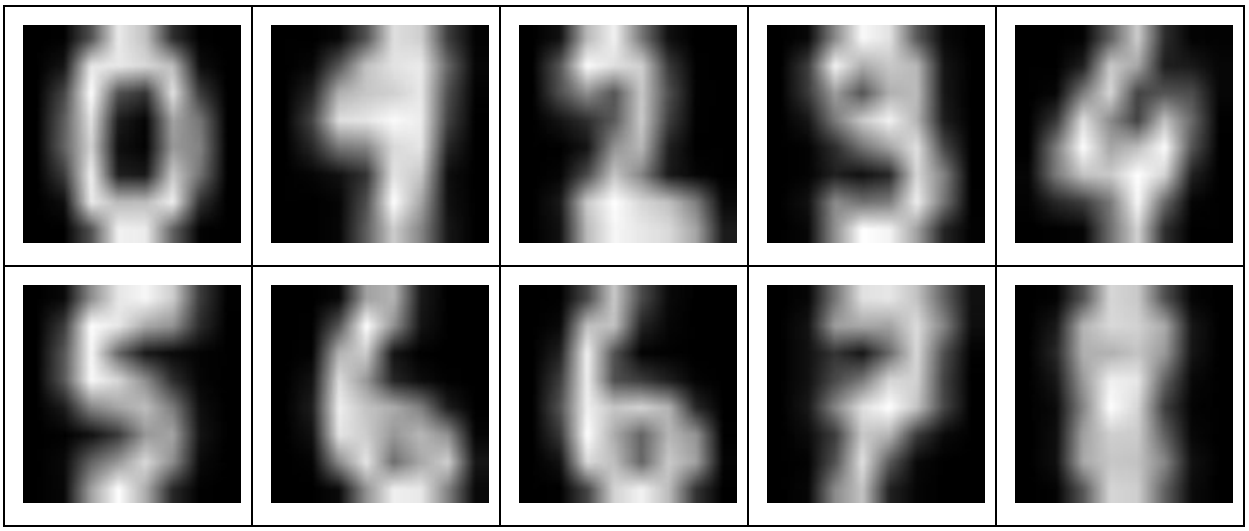
Average mean-square error: 644.52

Mean-square separation: 1324.85

Accuracy: 73.5%

Confusion Matrix:

		A	**c**	**t**	**u**	**a**	**l**				
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
P	**0**	176	0	1	0	0	0	1	0	0	0
r	**1**	0	61	2	0	4	0	1	0	7	3
e	**2**	0	21	150	0	0	0	0	0	1	0
d	**3**	0	1	8	165	0	26	0	0	31	145
i	**4**	2	0	0	0	144	1	1	0	0	0
c	**5**	0	0	0	1	0	152	0	0	2	3
t	**6**	0	4	0	0	0	1	176	0	2	0
e	**7**	0	0	3	9	26	2	0	176	10	27
d	**8**	0	95	13	8	7	0	2	3	121	2
	9	0	0	0	0	0	0	0	0	0	0



Experiment 2:

In the final experiment, we increased the number of centroids from ten to thirty. Instead of having to find the average of all possible ways of representing a digit, the classifier can find multiple sub-clusters that accurately characterize the digits. As a result, the accuracy of the classifier when using 30 centroids ranges between 89 - 94 percent and allows the classifier to cluster around more variation in the way the digits can be written.

On the following two pages is a summary of the results of experiment two along with the visualized clusters. You can see that the clustering is tighter by comparing the average mean square error of Experiment 1 with that of Experiment 2. By looking at the visualized centers on the last page, you can see that all of the ten classes are represented. There are a few clusters that show some ambiguity between digits, but for the most part, it’s clear which class corresponds to each centroid.

Average mean-square error: 473.28

Mean-square separation: 1552.37

Accuracy: 92.88%

Confusion Matrix:

		A	c	t	u	a	l				
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
P	**0**	177	0	0	0	0	1	0	0	0	1
r	**1**	0	177	4	0	9	0	0	0	19	1
e	**2**	0	0	160	2	0	0	0	0	0	0
d	**3**	0	0	1	166	0	0	0	0	2	5
i	**4**	1	0	0	0	168	1	0	1	0	7
c	**5**	0	0	0	2	0	174	2	0	2	2
t	**6**	0	0	0	0	1	0	177	0	1	0
e	**7**	0	0	2	3	0	0	0	168	1	0
d	**8**	0	2	10	5	3	0	2	3	141	3
	9	0	3	0	5	0	6	0	7	8	161