Tyler Sorg

Machine Learning

K-Means Project Report

Experiment 1: 10 Clusters

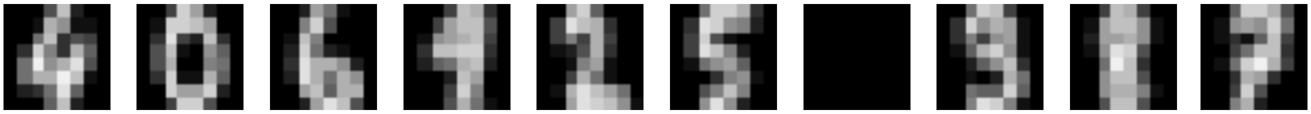
SSE: 2564781.86033 **SSS**: 74718.4871693 **Mean Entropy**: 0.10004797754 **Accuracy**: **73.51%**

Confusion Matrix for 10 Clusters

	P	R	E	D	I	C	T	E	D	
	176	0	0	0	2	0	0	0	0	0
A	0	56	21	1	0	0	4	0	100	0
C	1	2	149	9	0	0	0	3	13	0
T	0	0	0	165	0	1	0	9	8	0
U	0	5	0	0	162	0	0	6	8	0
A	0	1	0	31	1	148	1	0	0	0
L	1	0	0	0	1	0	176	0	3	0
	0	7	0	0	1	1	0	167	3	0
	1	8	1	34	0	4	2	2	122	0
	0	23	0	145	0	6	0	5	1	0

Note: The classes corresponding to the cluster number varies. What the matrix shows is how well the clustering classified each digit in order. As in, element (0,0) corresponds to how well the clustering classified actual 0s.

The mean entropy of each cluster was relatively low at 0.1. I do not know what good values of SSE and SSS are, so it is hard for me to comment on those. I did observe that maximizing the separation and minimizing the error usually led to lower entropies and higher accuracies. Overall, the classification accuracy was decent, especially compared to my multilayer neural network. Even if the clustering cannot classify a few digits at all, it can accurately classify most of the kinds of digits. The visualized clusters look like their associated digits when there are ten clusters. Here they are, ordered by cluster number:



Experiment 2: 30 Clusters

SSE: 1957201.06193 **SSS**: 891912.730607 **Mean Entropy**: 0.0145175007189 **Accuracy**: **89.37%**

Confusion Matrix for 30 Clusters

	P	R	E	D	I	C	T	E	D	
	177	0	0	0	1	0	0	0	0	0
A	0	149	20	1	0	0	4	0	5	3
C	0	4	168	0	0	0	0	2	3	0
T	0	0	2	152	0	3	0	3	7	16
U	0	6	0	0	172	0	0	0	3	0
A	0	0	0	1	1	152	1	0	18	9
L	2	3	0	0	1	2	173	0	0	0
	0	0	0	0	1	0	0	167	5	6
	0	22	2	3	2	0	1	1	136	7
	1	1	0	4	9	0	0	0	5	160

Here, the SSE is smaller, and the SSS is larger than in the first experiment. The mean entropy of the clustering is an order of magnitude smaller. The accuracy is noticeably higher, too. The cluster centers also look like their labels when there are 30 clusters.

Instead of putting them in the pdf, I will email the pictures in a zipped file. My program prints the labels corresponding to each cluster to a separate text file named experiment#_cluster_labels.txt.

Instructions for running my VERY SLOW (seriously, like 5 minutes) program:

I used the Pillow and NumPy libraries, so make sure you have those dependencies.

Put the optdigits files in their own folder called “optdigits” so that the python file and the optdigits folder are in the same path. Then just run the program using “`python k_means.py`” or however you normally run my programs. Thanks!