Portland State University

Deep Learning: Computational Structures and Programming

Winter 2021

Project #4

Author:
Hermann Yepdjio

Professor:
Dr. Suresh Singh

February 27, 2021

1 MNIST Data Set: K-means Clustering of 1,000 Images

Clustering Procedure:

I clustered the images using the k-means implementation in the scikit-learn library. Since the clustering algorithm assigns numbers randomly to each cluster, most clusters end up with the wrong label. To correct that, I first generated the confusion matrix using the initial numbers assigned to each cluster. Using this confusion matrix, I was able to assign each label to its corresponding cluster by looking in the confusion matrix, the cluster in which the label is the most represented. I followed the logic: if most 1s are in cluster #5, then cluster #5 should probably be labeled 1 instead of 5. Below is the final confusion matrix

Table 1: Confusion matrix.

	0	1	2	3	4	5	6	7	8	9
0	67	0	0	2	0	25	6	0	0	0
1	0	96	3	0	0	0	0	0	1	0
2	2	11	71	7	0	0	1	6	1	1
3	0	10	2	67	0	2	0	2	11	6
4	0	2	1	0	45	1	1	19	0	31
5	2	3	0	14	2	17	0	16	43	3
6	0	3	2	0	0	46	47	1	0	1
7	0	6	1	0	1	0	0	47	0	45
8	0	2	15	13	3	3	0	13	50	1
9	0	1	0	1	26	0	0	23	2	47

Accuracy = 55.4%. In the first assignment, we were able to achieve 88% classification accuracy Vs 55.4% for this approach. The difference is understandable since in the first assignment, we first showed the model multiple images for each class (supervised learning), while in this approach, we did not show any training data to the model (unsupervised learning). Instead, we just asked the model to find similarities among the testing data and classify them solely based on that. However, 55.4% shows that, this model is good as it was able to correctly classify more than half of the data without being given any information besides the images themselves.

2 K-means Clustering Using Feature Vectors

2.1 Without PCA

Table 2: Confusion matrix.

	0	1	2	3	4	5	6	7	8	9
0	785	1	5	49	8	54	70	0	24	4
1	0	524	1	0	1	3	2	2	2	465
2	4	34	698	35	26	19	27	2	23	132
3	4	8	67	667	16	41	12	13	59	113
4	2	36	4	0	572	23	12	289	3	59
5	13	6	1	361	30	428	26	14	72	49
6	17	25	6	4	32	17	838	0	2	59
7	0	35	3	0	259	4	0	599	0	100
8	9	53	9	132	22	42	15	10	591	117
9	9	11	2	10	474	3	1	419	1	70

Accuracy = 57.72%

The results of this approach are slightly better than those obtained in part 1. One reason to this improvement can be the number of samples. In this method, the clustering model has way more data to train on,and theoretically, should produce better results. Another reason is that, the data in this method first went through a trained model, which reduced the dimensionality of the images while keeping enough information that can be used to reconstruct the initial image. Therefore, the resulting images are just a compression of the original images and theoretically, if fed to a network, the results should be close to those obtained when using the original images.

2.2 With PCA (25 principal components)

Accuracy = 55.39% The results of this approach (using PCA) are slightly lower than those obtained with the previous method (Without using PCA). Such results were to be expected since, in theory, reducing the number of features normally should lead to a decrease in classification accuracy. However, this approach is still good because it allows to significantly reduce the processing time while having a very small impact on the final results.

Table 3: Confusion matrix.

	0	1	2	3	4	5	6	7	8	9
0	797	0	4	52	8	42	59	0	33	5
1	0	519	1	3	1	2	0	0	3	471
2	8	38	649	65	26	22	39	1	30	122
3	11	8	41	546	17	31	10	13	209	114
4	1	39	4	0	561	59	9	266	0	61
5	16	4	1	303	36	413	18	19	139	51
6	13	26	3	4	24	26	816	0	1	87
7	0	50	2	3	243	11	0	600	0	91
8	11	40	7	151	23	52	11	14	553	138
9	8	16	4	9	455	11	2	407	3	85