Luke Cathcart

CS 545

Dr. Doliotis

11/22/2019

# Homework 4

Experiment 1:

For this experiment, the clusters were randomly chosen. The seed for each was a random integer between 1 and 100. The actual seeds used were 2, 41, 37, 64, and 91.

The best run:

K = 10

Run: 5 AMSE: 666.89 MSS: 225.98 ME: 0.99 Seed: 91

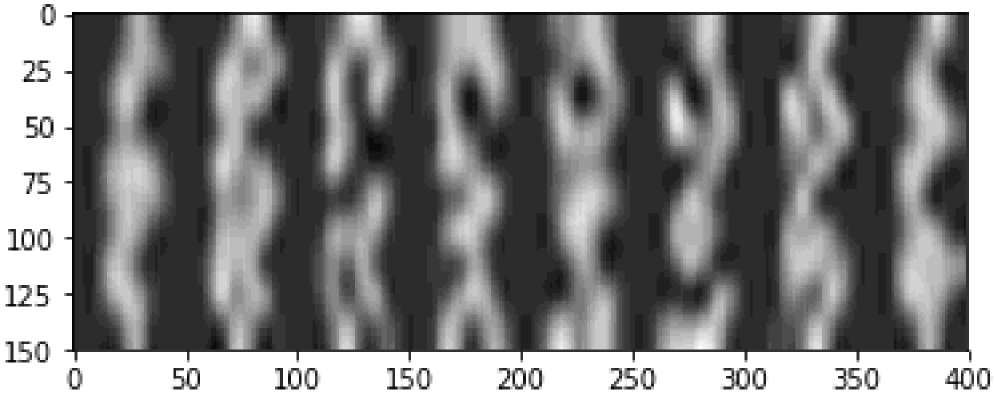
Classification Accuracy: 10.18%

Confusion Matrix:

```
[[ 29  48   0  34   0  66   0   0   1   0]
 [  0   0   0   3  24   0   0  31 121   3]
 [  0   1   0  82  37   0   2   2  48   5]
 [  1   0   0   2  25   0   5   1   1 148]
 [  0   0   1   4   0   0   0   5 171   0]
 [  1   0  58   0  67   0   0   0   1  55]
 [  1   0  27  69  27   0   0   0  57   0]
 [  0  87   0   6   7   0   0   0   3  76]
 [  0   0   1  27   6   0  47  61  27   5]
 [  0   0  93   2  69   1   2   1   3   9]]
```

Visualization Results:

Actual Cluster Centers: [1 5 1 5 5 0 5 8 2 3]



Discussion Paragraph:

K-Means did not perform well on this data. From both the accuracy and the visual results, the clusters don’t really look like the actual clusters that are listed above the picture, and the accuracy is barely above ten percent. There are a couple of reasons that this appears to be happening. First, the number associated with a cluster is not labeled 0-9 like the data is. This means that when this clustering is used on the test set, there a multiple digits that it can’t predict. Due to this, from this particular run, we are missing clusters associated with numbers 4, 6, 7, and 9. When we get the testing phase with our model, the model will never predict those numbers. I compared my results to the sklearn.cluster.KMeans, and found that the performance was about the same. Due to this, it appears that, when starting with random clusters, the ability of the K-Means to properly predict with 10 clusters was about as good as guessing.

Experiment 2:

For this experiment, the clusters were randomly chosen. The seed for each was a random integer between 1 and 100. The actual seeds used were 77, 82, 25, 48, and 8.

The best run:

K = 30

Run: 5 AMSE: 488.98 MSS: 110.08 ME: 0.5 Seed: 8

Classification Accuracy: 4.73%

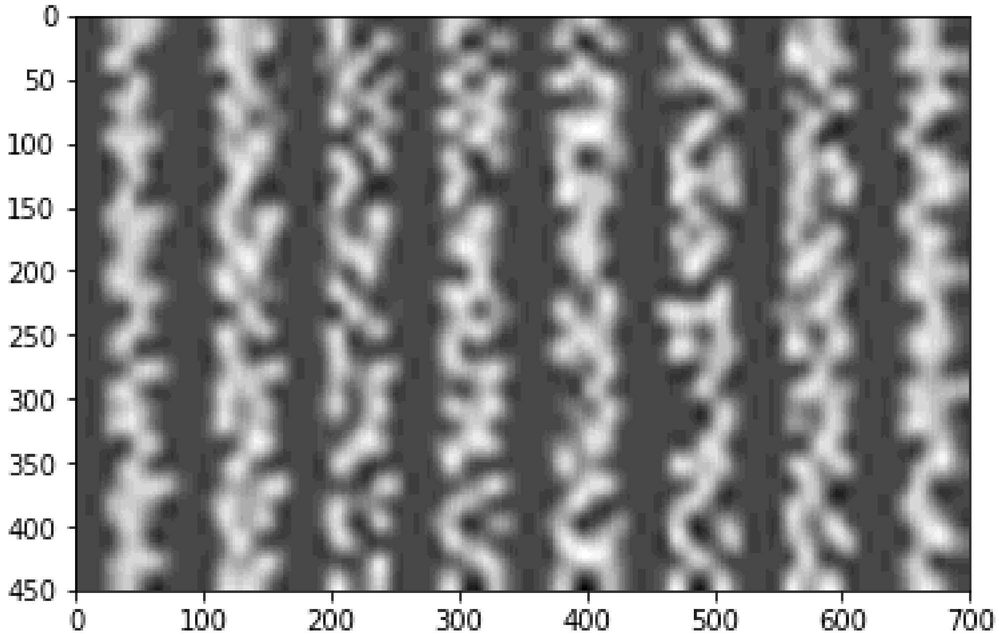
Confusion Matrix:

```
[[ 29   0   0   0  32  50  67   0   0   0   0]
 [  0   5   2   0   0  21   0   3 151   0   0]
 [  2  42   0   0  81  43   1   0   8   0   0]
 [ 32   2 136   0   0   3   5   0   3   2   0]
 [ 93   0   0   0   0   3  31   0   5   0  49]
 [  9   0   0   0   1   1   0   2 104  65   0]
 [  1   0   0   0  49  31   0  31  21  48   0]
 [ 53  75   0   0   0  47   0   0   2   0   2]
 [  0   3  45   0   0   2  66   4  50   2   2]
 [ 51  24   4   0   0   0   4  88   6   0   3]
 [  0   0   0   0   0   0   0   0   0   0   0]]
```

Visual Results:

Actual Clusters:

```
[5 5 0 0 4 0 4 7 4 0 6 8 4 9 6 2 8 1 1 7 2 8 5 5 8 6 9 1 0 8]
```



Discussion Paragraph:

With thirty clusters, K-Means is still not able to perform well. In fact, it performed worse than with ten clusters. However, we do find that the AMSE, MSS, and ME are all lower for the thirty clusters. However, we also see that the classification accuracy is lower, as well. We do see that the clusters have more of the class numbers present, but it is still missing 3. It appears that more clusters may reduce the error, entropy and separation of the clusters, but it does not make the classification better. Out of all of the clusters in the picture, there are a couple of sevens and eights that can be seen, but for the most part the clusters don't look like their pictures.