# Assignment #3: Unsupervised Learning and Dimensionality Reduction

Connor Bohan – connor.bohan@gmail.com

## Clustering

### Datasets

I chose to go with two, entirely new datasets for this project. The first dataset is a subset of the MNIST digit dataset. I chose 15 examples of digits zero through four which brings us to a total of 75 data. In the improbable case that you, the reader, are not familiar with the MNIST dataset, it contains a large quantity of handwritten, labeled pictures of digits (0-9). These pictures are 28 pixels by 28 pixels and are in greyscale. The second dataset I chose was the Stanford Dogs Dataset. This dataset contains a large quantity of pictures of dogs. I chose three dog species: Chow, Great Dane, and Irish Wolfhound and cropped 15 100x100 images of each species for a total of 45 data. I chose these dogs because they were different colors and should therefore be ‘far’ apart and in clusters. I also selected them because I like these breeds of dogs.

When loading these images, I scaled the RGB values for each pixel from the 0-255 range into the 0-1 range to help prevent the distance measurement from overflowing or becoming inaccurate. I then defined the distance between two images as the square root of the sum of squared difference for each color channel for each pixel in the image. If that doesn’t make much sense, I pretended that the images were a single point in a high dimensional space (for the 100x100 images it would be 30,000-dimensional space or 100 for the width x 100 for the height x 3 for the color channels) and then found the Euclidean distance between those two points. Why did I pick this metric for distance? I picked it because that was the metric I was taught in my computer vision course and it worked well there.

The MNIST dataset is interesting because it should be relatively easy for clustering to work well on. Digits are distinct from one another and different instances of the same digit should have a large amount of overlap. This data can be used to provide a baseline of how well my clustering algorithms perform. The dog dataset on the other hand is a hard problem. Separate instances with the same label will not necessarily look similar. You can think of the dog dataset as being three gaussian distributions with very large sigma values. On the other hand, the MNIST dataset is five gaussian distributions with very small sigma values. The dog dataset is there to test the limits of what my clustering implementations can achieve.

### K-Means Clustering

For the MNIST dataset, I set k equal to five since there are five different digits in my subset. K-Means performed adequately but tended to converge on suboptimal solutions. The biggest problem by far was that it tended to create clusters that had many individuals with several labels while creating other clusters that had few individuals. The best possible clustering for the MNIST dataset would be five clusters each containing all the individuals with one type of label (i.e. cluster one would contain all images of a zero, cluster two would contain all images of a one, etc.). I would say this clustering has zero error. A clustering like this, except with one individual moved to a different cluster, would have one error. My metric for error is: for each individual in a cluster that isn’t a member of the majority, the error goes up by one. This metric allows multiple clusters to be cluster of the same label (e.g. there could be two clusters that contain more instances of the digit four than anything else) but it was the best metric I could think of.

With our error metric defined, here are the results of several test runs of K-means clustering on the MNIST dataset (note that there 75 individuals in the dataset):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test Run | 1 | 2 | 3 | 4 | 5 | 6 |
| Error | 25 | 32 | 23 | 33 | 27 | 14 |
| % Error | 33.3% | 42.6% | 30.6% | 44% | 36% | 18.6% |

Figure 1: The number of individuals placed in the wrong cluster over six runs of K-Means clustering on the MNIST dataset. The ‘% Error’ row is just the Error row divided by the number of individuals in the dataset (75).

Both test runs two and four had ‘super clusters’ that contained almost all individuals of several labels. For example: if a cluster had all 15 members of the one label and 14 members of the two label, then all 14 members of the two label were counted as being in the wrong cluster. As you can see, K-means clustering did much better than randomly assigning the images to clusters. However, it generally didn’t do well enough to be useful. It certainly won’t be replacing neural nets for digit recognition any time soon.

For the dog dataset, I set k equal to three since there are three breeds of dog represented in the data. The following table contains the results of several test runs of K-means clustering on the dog dataset. The dog dataset contains only 45 individuals so if you compare this table to the MNIST table, make sure to use the ‘% Error’ row.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test Run | 1 | 2 | 3 | 4 | 5 | 6 |
| Error | 16 | 11 | 17 | 17 | 20 | 18 |
| % Error | 35.5% | 24.4% | 37.7% | 37.7% | 44.4% | 40% |

Figure 2: The number of individuals placed in the wrong cluster over six runs of K-Means clustering on the dog dataset. The ‘% Error’ row is just the Error row divided by the number of individuals in the dataset (45).

K-means performed very similarly on both datasets, which I find surprising. The dog dataset seems to be the much noisier of the two. I expected the results on this dataset to be much worse than they were. This discrepancy could be because there are fewer individuals in the dataset, or because there were fewer clusters.

### Expectation Maximization Clustering

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test Run | 1 | 2 | 3 | 4 | 5 | 6 |
| Error | 25 | 32 | 23 | 33 | 27 | 14 |
| % Error | 33.3% | 42.6% | 30.6% | 44% | 36% | 18.6% |

Figure 3: The number of individuals placed in the wrong cluster over six runs of Expectation Maximization clustering on the MNIST dataset. The ‘% Error’ row is just the Error row divided by the number of individuals in the dataset (75).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test Run | 1 | 2 | 3 | 4 | 5 | 6 |
| Error | 16 | 11 | 17 | 17 | 20 | 18 |
| % Error | 35.5% | 24.4% | 37.7% | 37.7% | 44.4% | 40% |

Figure 4: The number of individuals placed in the wrong cluster over six runs of Expectation Maximization clustering on the dog dataset. The ‘% Error’ row is just the Error row divided by the number of individuals in the dataset (45).