Katie Hughes

Jasper LaFortune

Assignment 5

For this assignment, we implemented k-means clustering on a dataset of handwritten digits ‘4’ and ‘9’ from the USPS. We performed clustering with 2, 4, 6, and 8 clusters on both the original dataset and a reduced-dimensional dataset computed by Principal Component Analysis (PCA):

```
Original Data Confusion Matrix for k = 2 (purity = 0.542143)
   307   366
   393   334

Reduced Dimensional Data Confusion Matrix for k = 2 (purity = 0.768571)
   170   546
   530   154

---------------------------------------------

Original Data Confusion Matrix for k = 4 (purity = 0.420714)
   151   285
   391    17
    72   182
    86   216

Reduced Dimensional Data Confusion Matrix for k = 4 (purity = 0.405714)
   170   212
   343    39
    77   209
   110   240

---------------------------------------------

Original Data Confusion Matrix for k = 6 (purity = 0.362857)
    38   205
   176    29
    58   175
   230    29
    43   169
   155    93

Reduced Dimensional Data Confusion Matrix for k = 6 (purity = 0.474286)
```

```
   201     2
    35   180
    59   109
   170    57
    79   202
   156   150

--------------------------------------------

Original Data Confusion Matrix for k = 8 (purity = 0.320714)
    18   150
   157    12
    43   170
   114    83
    41   164
    28    74
   137    16
   162    31

Reduced Dimensional Data Confusion Matrix for k = 8 (purity = 0.397143)
    29   121
   184     1
    12   181
    12   153
    89    86
   132    48
    88    84
   154    26
```

In general, class purity decreases as the number of clusters, k, increases. This makes sense because there are two true class labels, so we would expect two true clusters. Additional clusters increase the chance for overlap between cluster labels and class labels, decreasing the purity. Additionally, the clustering on the reduced dimensional dataset had generally higher purity than the clustering on the original dataset. This is because Euclidean distance is overly sensitive in high dimensional spaces. That is, if two data points are very different in one dimension but identical in the rest, they will be considered “farther apart” than two data points a medium distance apart in all dimensions. PCA resolves this issue by reducing the number of dimensions in which this type of error can occur. Clustering also ran faster on the reduced-dimensional dataset because computing distance between points is a O(d) operation, where d is the number of the dimensions of the points.