

---

## Table of Contents

.....	1
In this script, you need to implement three functions as part of the k-means algorithm. ....	1
Initialize Data Set .....	1
After initializing, you will have the following variables in your workspace: .....	2
To visualize an image, you need to reshape it from a 784 dimensional array into a 28 x 28 array. ....	2
After importing, the array 'train' consists of 1500 rows and 785 columns. ....	3
This next section of code calls the three functions you are asked to specify .....	3
The next line initializes the centroids. Look at the initialize_centroids() .....	3
Initialize an array that will store k-means cost at each iteration .....	4
This for-loop enacts the k-means algorithm .....	4
This section of code plots the k-means cost as a function of the number .....	4
This next section of code will make a plot of all of the centroids .....	5
Function to initialize the centroids .....	6
Function to pick the Closest Centroid using norm/distance .....	6
Function to compute new centroids using the mean of the vectors currently assigned to the centroid. ....	7

```
clear all;  
close all;
```

## In this script, you need to implement three functions as part of the k-means algorithm.

These steps will be repeated until the algorithm converges: 1. initialize\_centroids This function sets the initial values of the centroids 2. assign\_vector\_to\_centroid This goes through the collection of all vectors and assigns them to centroid based on norm/distance 3. update\_centroids This function updates the location of the centroids based on the collection of vectors (handwritten digits) that have been assigned to that centroid.

## Initialize Data Set

These next lines of code read in two sets of MNIST digits that will be used for training and testing respectively.

```
% training set (1500 images)  
train=csvread('mnist_train_1500.csv');  
trainsetlabels = train(:,785);  
train=train(:,1:784);  
train(:,785)=zeros(1500,1);  
  
% testing set (200 images with 11 outliers)  
test=csvread('mnist_test_200_woutliers.csv');  
% store the correct test labels  
correctlabels = test(:,785);  
test=test(:,1:784);
```

---

```
% now, zero out the labels in "test" so that you can use this to
assign
% your own predictions and evaluate against "correctlabels"
% in the 'csl_mnist_evaluate_test_set.m' script
test(:,785)=zeros(200,1);
```

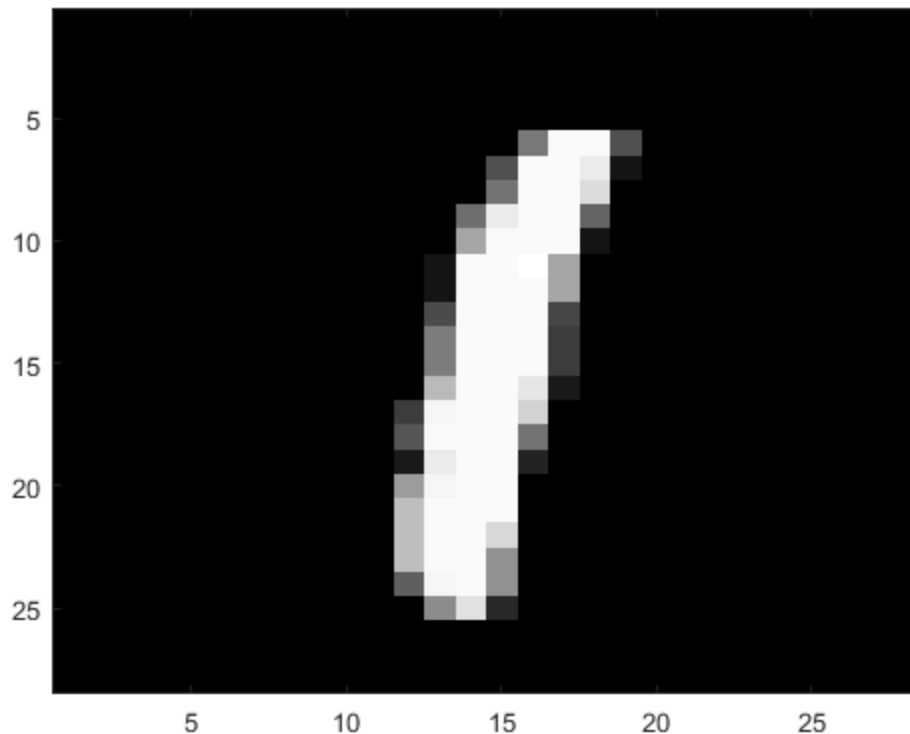
## After initializing, you will have the following variables in your workspace:

1. train (a 1500 x 785 array, contains the 1500 training images) 2. test (a 200 x 785 array, containing the 200 testing images) 3. correctlabels (a 200 x 1 array containing the correct labels (numerical meaning) of the 200 test images)

## To visualize an image, you need to reshape it from a 784 dimensional array into a 28 x 28 array.

to do this, you need to use the reshape command, along with the transpose operation. For example, the following lines plot the first test image

```
figure;
colormap('gray'); % this tells MATLAB to depict the image in grayscale
testimage = reshape(test(1,[1:784]), [28 28]);
% we are reshaping the first row of 'test', columns 1-784 (since the
785th
% column is going to be used for storing the centroid assignment.
imagesc(testimage); % this command plots an array as an image. Type
'help imagesc' to learn more.
```



**After importing, the array 'train' consists of 1500 rows and 785 columns.**

Each row corresponds to a different handwritten digit ( $28 \times 28 = 784$ ) plus the last column, which is used to index that row (i.e., label which cluster it belongs to. Initially, this last column is set to all zeros, since there are no clusters yet established.

**This next section of code calls the three functions you are asked to specify**

```
k = 16; % set k
max_iter = 10; % set the number of iterations of the algorithm
```

**The next line initializes the centroids. Look at the `initialize_centroids()`**

function, which is specified further down this file.

```
centroids=initialize_centroids(train,k);
```

---

## Initialize an array that will store k-means cost at each iteration

```
cost_iteration = zeros(max_iter, 1);
```

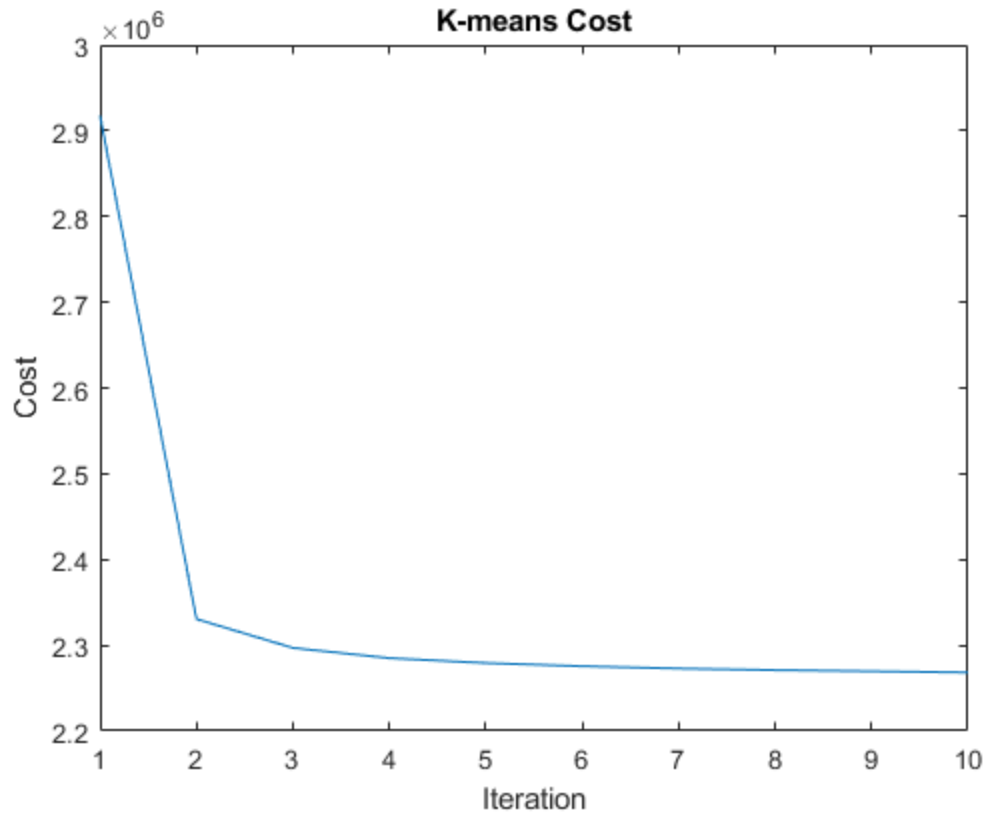
## This for-loop enacts the k-means algorithm

```
%run the algorithm for the desired number of iterations
for iter = 1:max_iter
    %iterate for each image
    for image = 1:length(train)
        %grab the distance of each vector (excluding the last element)from
        its
        %respective centroid using vec_distance function.
        [index, vec_distance] = assign_vector_to_centroid(train(image,
        1:784) ,centroids(:, 1:784));
        %assign each image vector to its centroid by putting its centroid
        index
        %in the 785th column
        train(image, 785) = index;
        %sum the cost for each iteration and assign it to its spot in
        the
        %cost_iteration vector
        cost_iteration(iter, 1) = cost_iteration(iter, 1) +
        vec_distance;
    end
    %use update_centroids to get new centroids
    centroids = update_Centroids(train, k);
end
```

## This section of code plots the k-means cost as a function of the number

of iterations

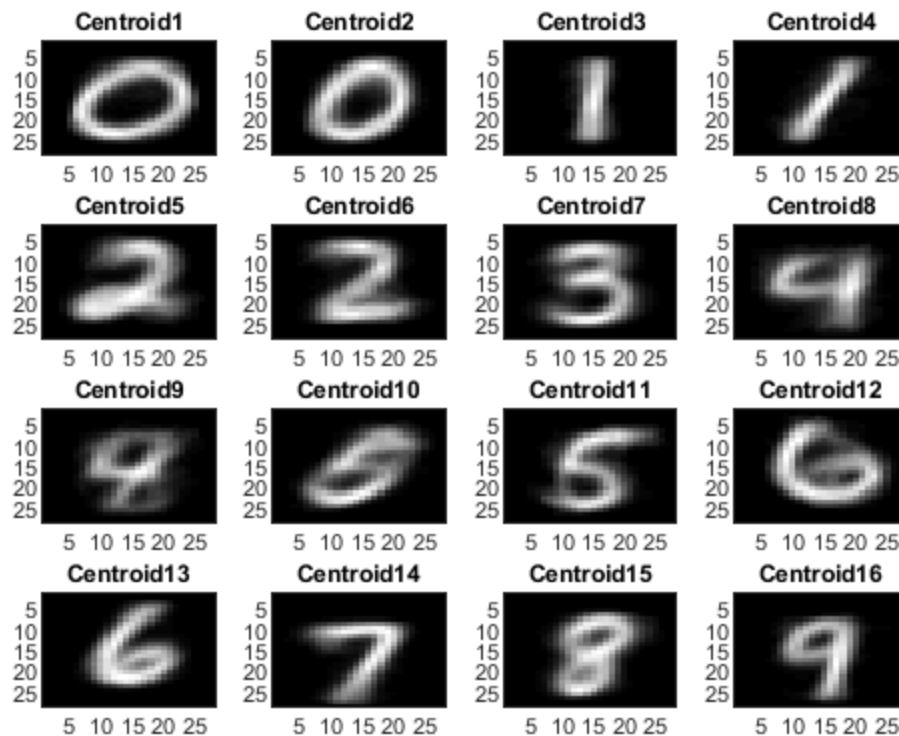
```
%create a plot of the cost_iteration vector, label axes, and make
    title.
figure;
plot(cost_iteration);
xlabel("Iteration");
ylabel("Cost");
title("K-means Cost");
```



**This next section of code will make a plot of all of the centroids**

Again, use help [functionname](#) to learn about the different functions that are being used here.

```
figure;  
colormap('gray');  
  
plotsize = ceil(sqrt(k));  
  
for ind=1:k  
  
    centr=centroids(ind,[1:784]);  
    subplot(plotsize,plotsize,ind);  
  
    imagesc(reshape(centr,[28 28]));  
    title(strcat('Centroid ',num2str(ind)))  
  
end
```



## Function to initialize the centroids

This function randomly chooses  $k$  vectors from our training set and uses them to be our initial centroids. There are other ways you might initialize centroids. \*Feel free to experiment.\* Note that this function takes two inputs and emits one output ( $y$ ).

```
function y=initialize_centroids(data,num_centroids)
%choice of initial centroids
choice_initial = [5, 33, 12, 24, 31, 27, 7, 116, 147, 50, 68, 93, 19,
    131, 20, 170];
centroids=data(choice_initial,:);
y=centroids;
end
```

## Function to pick the Closest Centroid using norm/distance

This function takes two arguments, a vector and a set of centroids. It returns the index of the assigned centroid and the distance between the vector and the assigned centroid.

```
function [index, vec_distance] =
    assign_vector_to_centroid(data,centroids)
    %find distance between the data vector and each centroid
    dist = pdist2(data, centroids, "euclidean", "Smallest", 1);
```

---

```
    %find the centroid closest to the data vector and find that
    distance.
    [vec_distance, index] = mink(dist, 1);
end
```

## Function to compute new centroids using the mean of the vectors currently assigned to the centroid.

This function takes the set of training images and the value of k. It returns a new set of centroids based on the current assignment of the training images.

```
function new_centroids=update_Centroids(data, K)

%create a K by 785 matrix filled with zeros
new_centroids = zeros(K, 785);

%iterate for each centroid
for cents = 1:K
    %grab each image vector mapped to the centroid
    images_by_cent = data(:, 785) == cents;
    %use indexing to access each of those vectors
    vector_by_cent = data(images_by_cent, :);
    %average those vectors to create the new centroid
    new_centroids(cents, :) = mean(vector_by_cent);
end
end
```

*Published with MATLAB® R2021a*