In this project I implemented K-means clustering algorithm to identify main clusters in the data and use the discovered centroid of cluster for classification. Specifically,

- Implemented K-means clustering algorithm to identify clusters in a two-dimensional toy-dataset.
- Implemented image compression using K-means clustering algorithm.
- Implemented classification using the centroids identified by clustering on digits dataset.
- Implemented K-means++ clustering algorithm to identify clusters in a two-dimensional toy-dataset i.e. implement the K-means++ function to compute the centers.

Dataset for K-Means Clustering

We will use 2 datasets - 2-D Toy Dataset and Digits datasets for K means part.

Toy Dataset is a two-dimensional dataset generated from 4 Gaussian distributions. We will use this dataset to visualize the results of our algorithm in two dimensions.

We will use digits dataset from sklearn to test K-means based classifier and generate digits using Gaussian Mixture model. Each data point is an 8 × 8 image of a digit. This is similar to MNIST but less complex. There are 10 classes in digits dataset.

Link for Digits dataset: sklearn. datasets. digits
http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits

After successful implementation of K-means, K-means++ and Image compression algorithms:

- I was able to compare the clusters identified by the algorithm against the real clusters for toy dataset.
- Able to use K-means and K-means++ as a classification algorithm with the prediction accuracies of 77% and 72% on Digits dataset.
- Used K-means algorithm for image compression, to generate a compressed image with mean-squared error close to 0.0098.