CSE5334: Data Mining
Assignment 1

Problem 1: K-means Clustering

- This problem is about a technique of Unsupervised Learning called K-means Clustering.
- The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset.
- It accomplishes this using a simple concept of what the optimal clustering looks like:

1. The "cluster center" is the arithmetic mean of all the points belonging to the cluster.
2. Each point is closer to its own cluster center than to other cluster centers.

Algorithm:

Given an initial set of k means, the algorithm proceeds by alternating between two steps:

- **Assignment step:** Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.
- **Update step:** Calculate the new means to be the centroids of the observations in the new clusters.

  The algorithm has converged when the assignments no longer change. The algorithm does not guarantee to find the optimum.

Implementation:

Step 1: Generated two sets of 2-D Gaussian random data, each set containing 500 samples and merged the data into a single dataset.

Step 2: Chosen the initial centroids from this dataset based on the k values in the function

mykmeans(X, k, c).

Step 3: Found the Euclidean distance of all the data points from these initial centroids, collected minimum value of distance for each data point and assigned them to dictionary (Key-value) fashion.

Step 4: Calculated the new mean for the newly created clusters. These new means are the updated centroids.

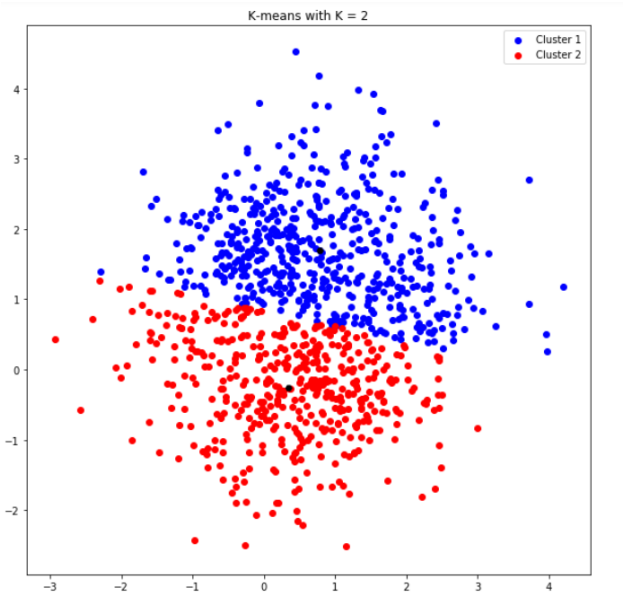
Step 5: Repeated above mentioned steps till the distance between initial centroids and updated centroids converges to less than 0.001.

Step 6: function mykmeans returns dictionary with clustered data points, number of times, the Assign-Update steps performed and centroids of each cluster.

CSE5334: Data Mining
Assignment 1

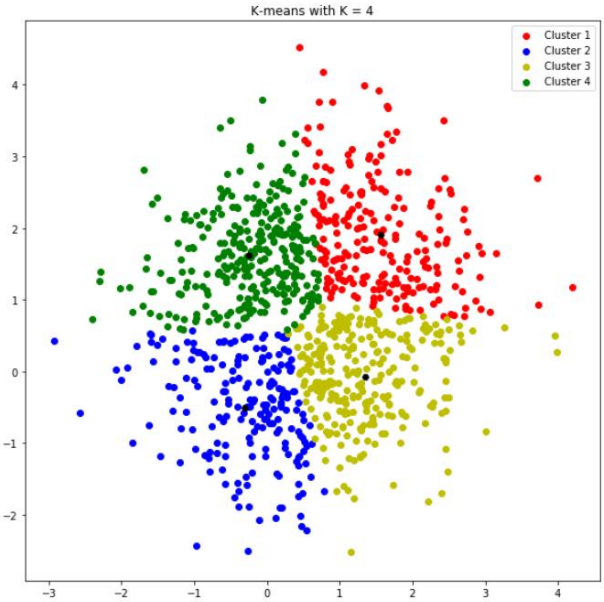
Problem 1.2

Scatter plot of the data and the centers of clusters found after plugging data generated with k = 2 and initial centers c1 = (10, 10) and c2 = (-10, -10) into function **cluster = mykmeans(X, k, c).**



Problem 1.3

Scatter plot of the data and the centers of clusters found after plugging data generated with k = 4 and initial centers c1 = (10, 10), c2 = (-10, -10), c3 = (10, -10) and c4 = (-10, 10) into function
cluster = mykmeans(X, k, c).



CSE5334: Data Mining
Assignment 1

Problem 2: Non-parametric Density Estimation

- This problem is about a technique of Unsupervised Learning called Non-Parametric Kernel Density Estimation.
- Kernel density estimation is a non-parametric method of estimating the probability density function (PDF) of a continuous random variable. It is non-parametric because it does not assume any underlying distribution for the variable.

Algorithm:

- Let (x1, x2, …, xn) be a univariate independent and identically distributed sample drawn from some distribution with an unknown density ƒ.
- To estimate the shape of this function f, kernel density estimator of f can be found as below:

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

where K is the kernel — a non-negative function.

h > 0 is a smoothing parameter called the bandwidth.

Implementation:

Step 1: Generated N = 1000 Gaussian random data with mean = 5 and std = 1. This data has been given as an input to the **function [p, x] = mykde (X, h)** with h = {.1, 1, 5, 10}.

Step 2: A sample of randomly chosen data points is used to find values of Kernel function for each points and summation of these discrete values over N data points of sample.

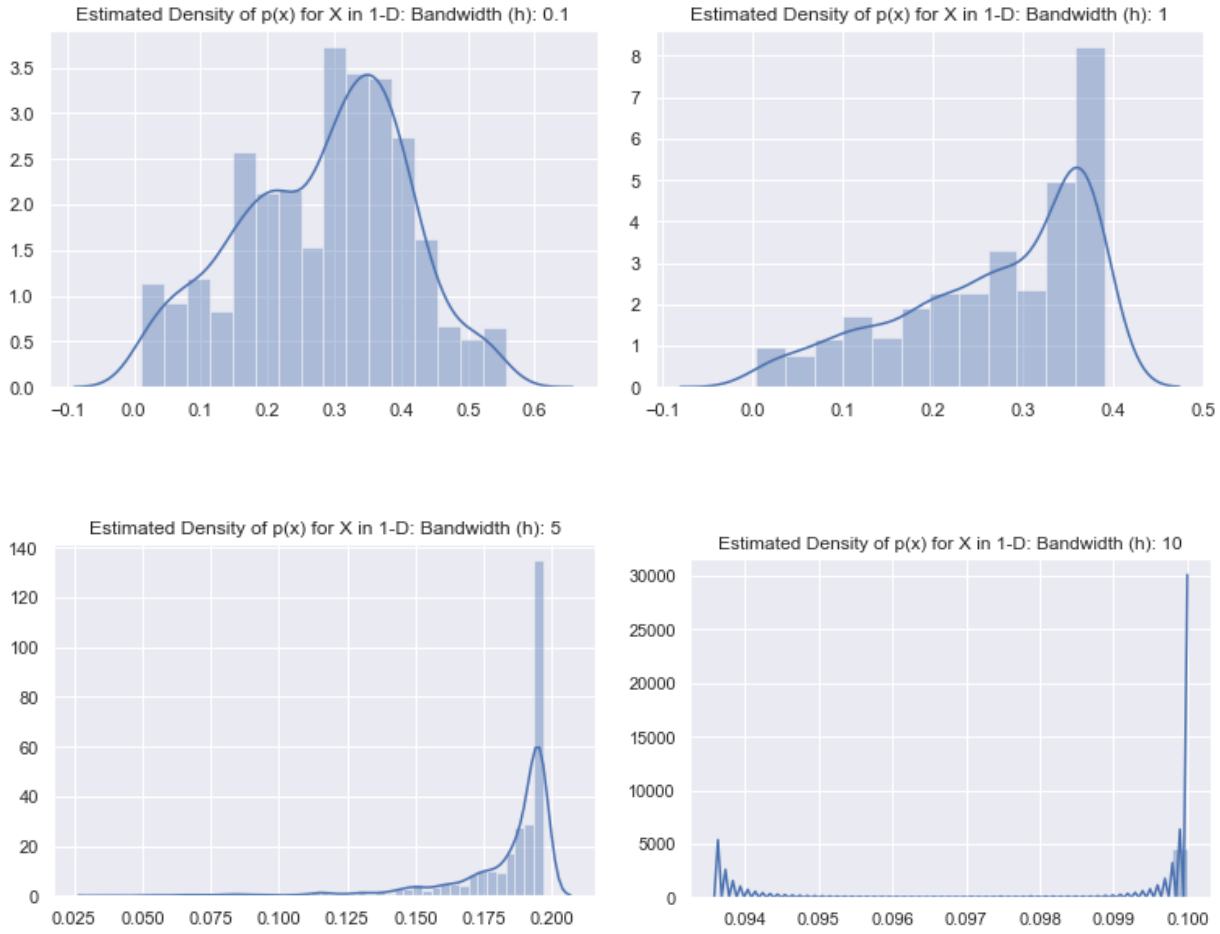
Step 3: Based on the value of the kernel function, parzen window is applied to each kernel function values for each point.

Step 4: This way for a pre-defined bandwidth, probability density of each data-point in a sample is calculated and this process is repeated for every bandwidth value.

CSE5334: Data Mining
Assignment 1

Problem 2.2

Histogram of X along with the figures of estimated densities after plugging data generated as N = 1000 Gaussian random data with mean = 5 and std = 1 into **function [p, x] = mykde(X,h)** with h = {.1, 1, 5, 10}.



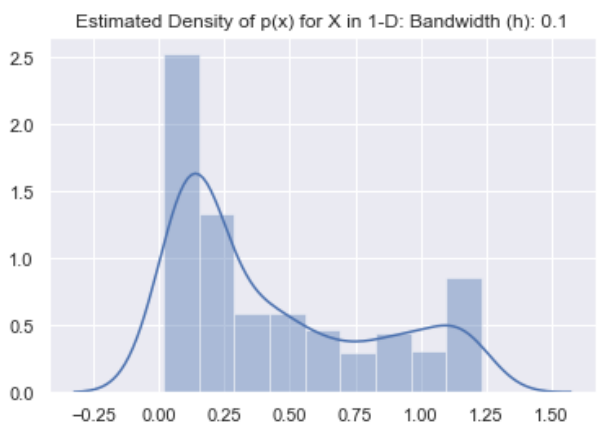
Bandwidth selection

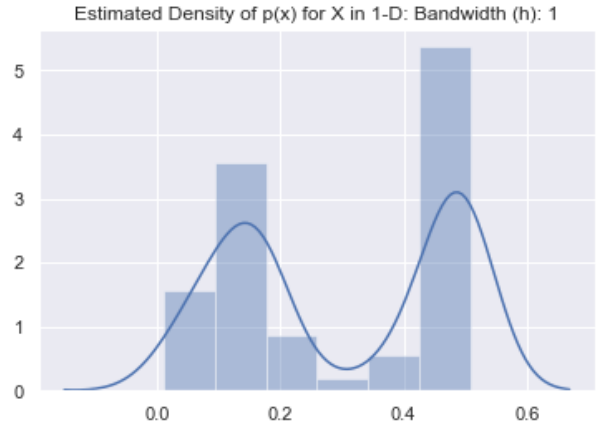
- The bandwidth of the kernel is a free parameter which exhibits a strong influence on the resulting estimate.
- The Density Estimation curve with a bandwidth of h = 0.1 is optimally smoothed since its density estimate is close to the true density. As the value of bandwidth increases, the curve is getting oversmoothed since using the bandwidth h = 1, 5 & 10 obscures much of the underlying structure.

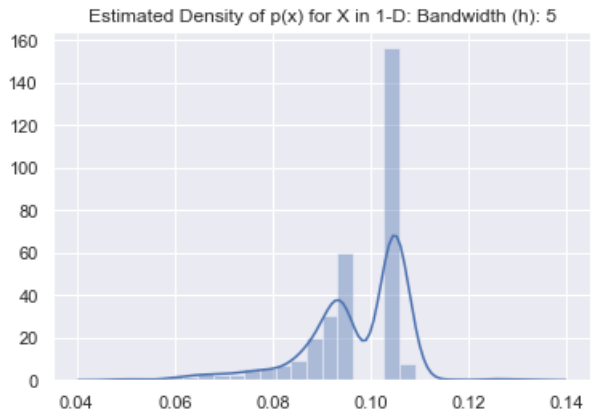
CSE5334: Data Mining
Assignment 1

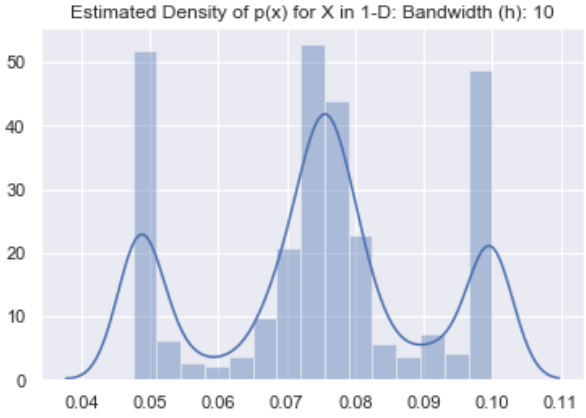
Problem 2.3

Histogram of X along with the figures of estimated densities after plugging data generated by merging two N = 1000 Gaussian random data with (mean = 5 and std = 1) and with (mean = 0 and std = 0.2) into **function [p, x] = mykde(X,h)** with h = {.1, 1, 5, 10}.









Bandwidth selection

- The bandwidth of the kernel is a free parameter which exhibits a strong influence on the resulting estimate.
- The Density Estimation curve with a bandwidth of h = 1 is optimally smoothed since its density estimate is close to the true density. As the value of bandwidth increases, the curve is getting oversmoothed since using the bandwidth h = 5 & 10 obscures much of the underlying structure. In comparison, the curve is under-smoothed since it contains too many spurious data artifacts arising from using a bandwidth h = 0.1, which is too small.

CSE5334: Data Mining
Assignment 1

Problem 2.4

Histogram of estimated densities of 2 sets of 2-D Gaussian random data after plugging data into **function [p, x] = mykde(X,h)** with h = {.1, 1, 5, 10}.