As part of this assignment, you will write the core parts of K-means algorithm. We will assume that the input is a DxN matrix (i.e., N D-dimension points), and the initial centers are represented by a DxK matrix (i.e., K D-dimensional points).

Task 1: Generate randomized input data (4 points)

Write a R function called “genData(N,D,K)” which returns a single R list with two components “all_points” and “centers”:

- “all_points” is a DxN matrix. Elements in “all_points” are selected uniformly at random from the range 1 to D*N (with replacement)
- “centers” is a DxK matrix. Elements in “centers” are K of the input points from “all_points” (i.e., D-dimensional points) selected uniformly at random (but without replacement).

Task 2: Calculate distance to closest center (15 points)

Write a R function called “closestCenters(all_points, centers)” which returns a single R list with two components “new_centers” (DxN matrix) and “label” (1xN matrix).

- For each element in “all_points” find the closest point in “centers”, and store the center id in label. For example, if the first point in “all_points” is closest to center number three then label[1] will be 3. Use Euclidean distance measure. To obtain full points, you should use matrix multiplication and subtraction to calculate distances between points and centers (otherwise calculations may be too slow in R).
- “new_centers” is a DxK matrix. It is obtained by averaging the points that have the same label.

Task 3: Implement K-means algorithm (6 points)

Write a R function called “myKmeans(all_points, centers, niter)” which takes the arguments “all_points” (DxN), initial center (DxK), and niter (number of iterations), and return a single list with two components “new_centers” and “label”. The “myKmeans” function should successively refine the new cluster centers by calling the “closentCenters” function (developed in Task 2) for “niter” iterations.

Curious about correctness of your code?

R provides a built-in “kmeans” function. You can use the kmeans function (with algorithm="Lloyd") to check the accuracy of the centers you have calculated. Note that there may be some differences between your results and what R’s kmeans returns [There are no points for this step].