

# MSMS 206 : Practical 06

Ananda Biswas

April 19, 2025



**Question :** Use  $k$ -means clustering to divide *iris* dataset into 3 clusters.

⊕ After a choice of initial centroids, the  $k$ -means clustering algorithm is as follows :

- (1) calculate the distance of each data-point from each of the centroids
- (2) assign each of the data-points to its closest centroid
- (3) relocate the centroids to the average location of the data-points of similar group

And we repeat this procedure until the assignments don't change after the centroid locations were recomputed.

```
df <- iris[, -5]
```

```
dim(df)
```

```
## [1] 150 4
```

```
m <- dim(df)[1] # number of data-points  
n <- dim(df)[2] # dimension of data-points  
  
k <- 3 # number of clusters
```

```
X <- as.matrix(df)
```

Now we initialize the centroids as 3 randomly chosen data-points.

```
random_index <- sample(m, k)  
  
centroid <- X[random_index, ]
```

We now deploy our  $k$ -means clustering algorithm.

```
cluster <- c()  
  
repeat{  
  dist_mat <- matrix(0, nrow = m, ncol = k)
```

```

for (i in 1:k) {
  d <- apply(X, 1, FUN = function(x) return(x - centroid[i, ]))

  d <- matrix(d, nrow = m, ncol = n, byrow = TRUE)

  dist_mat[,i] <- sqrt(diag( d %*% t(d) ) )
}


cluster <- apply(dist_mat, 1, FUN = function(x) return(which(x == min(x))[1]))

new_centroid <- matrix(data = 0, nrow = k, ncol = n)

for (i in 1:k) {
  new_centroid[i, ] <- mean(X[which(cluster == i), ])
}

if(any(centroid - new_centroid != 0)){
  centroid <- new_centroid
} else{
  break
}
}

```

 The final clustering of the data-points is as follows :

```

cluster

##      [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##      [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 2 1 1 1 1 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      [75] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 3 1 2 1 2 2 2 2 1 2 2 2 2
##     [112] 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 1 2 2 2 1 2 2 2 1 2 2 2 1 2
##     [149] 2 1

```

In *iris* dataset, frequency distribution of 3 species was :

```

table(iris[,5])

##
##      setosa versicolor  virginica
##          50          50          50

```

Our *k*-means algorithm categorizes the *iris* dataset with the frequency distribution as follows :

```

table(cluster)

## cluster
##      1  2  3
##     58 38 54

```