10.41 K-means as an approximate matrix factorization

Blake Baird

2022-03-30

N n-vectors as input vectors: $X = \{x_1, x_2, ..., x_N\}$
Run k-means to get k group representatives (also n-vectors): $Z = \{z_1, z_2, ..., z_k\}$
Encode the assignment of vectors to groups in the **clustering matrix**,C. $C_{ij} = 1$ if vector x_j belongs group i.

Let N = 3 and k = 2. An example of C when x_1 and x_3 belong to group 2 and x_2 belongs to group 1:

```
matrix(c(0,1,0,1,0,1),ncol=3,byrow=TRUE)
```

```
##      [,1] [,2] [,3]
## [1,]    0    1    0
## [2,]    1    0    1
```

Q1. Give an interpretation of the columns of of the matrix $X - ZC$ and the squared norm $||X - ZC||^2$.

Each column of X-ZC is the difference between the entries of the x's and their nearest group representative. The squared norm is the mean squared distance of the x's from their group representatives.

$J_{clust} = ||X - ZC||^2/N$
$||X - ZC||^2 = J_{clust}N$

Q2. Justify the statement that the goal of K-means is to find the best Z and C to minimize J_{clust}.

Minimizing the squared norm is the same as minimizing J_{clust} over all possible Zs and Cs.

$X = QR$ is one form of a factorization And $X \approx ZC$ is an approximate form of matrix factorization as well.