Clustering

Unsupervised learning; group data points which are similar to each other

$\mu_1, \mu_2, ..., \mu_K$ cluster centers

We minimize the total squared distance from data points to cluster centers:

$J(c, \mu) = \Sigma_{i=1}^{N} ||x^{(i)} - \mu_{c^{(i)}}||^2$

K-means Clustering Algorithm:

Randomly initialize $\mu_1, \mu_2, ..., \mu_K$

Repeat until convergence:

for each i:

$c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||^2$

for each $j \in \{1, ..., K\}$:

$\mu_j = \frac{\Sigma_{i=1}^{N} 1\{c^{(i)} = j\} x^{(i)}}{\Sigma_{i=1}^{N} 1\{c^{(i)} = j\}}$

Random intialization

The algorithm may converge to local optimums.

We can run the algorithm with random initialization for 10 times, and choose the one with lowest $J(c, \mu)$.

$\mu_1 = x^{(j)}$ // randomly initialize first centroid as the jth data point

for k'' = 2 to k:

$d_j = \min_{k' < k''} ||x^{(j)} - \mu_{k'}||, \forall j$

$p_j = \frac{d_j^2}{\Sigma_{i=1}^{m} d_j^2}, \forall j$

j = randomly chosen with probability p_j

$\mu_{k''} = x^{(j)}$

run k-means using μ as initial centers

Documents as feature vectors

Term frequency - Inverse Document Frequency