CS3244 Tutorial 5

Zhuang Jianming
A0214561M
T13

CL1

a) Iteration 1 :

ID	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
cluster centroid	C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2

$$\text{new } C_1 = \left(\frac{1+2+2+4+5}{5}, \frac{2+5+10+9+8}{5}\right) = (2.8, 6.8)$$

$$\text{new } C_2 = \left(\frac{6+7+8}{3}, \frac{4+5+4}{3}\right) = \left(7, \frac{13}{3}\right)$$

Iteration 2 :

ID	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
cluster centroid	C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2

cluster membership does not change
$\Rightarrow$ centroids do not change

final cluster centroids $\Rightarrow$ $C_1 = \underline{(2.8, 6.8)}$, $C_2 = \underline{\left(7, \frac{13}{3}\right)}$

b)

$$SSE = \sum_{i=1}^{2} \sum_{x \in C_i} dist^2(c_i, x)$$

$$= (1-2.8)^2 + (2-6.8)^2 + (2-2.8)^2 + (5-6.8)^2 + (2-2.8)^2 + (10-6.8)^2$$
$$+ (4-2.8)^2 + (9-6.8)^2 + (5-2.8)^2 + (8-6.8)^2 + (6-7)^2 + \left(4-\frac{13}{3}\right)^2$$
$$+ (7-7)^2 + \left(5-\frac{13}{3}\right)^2 + (8-7)^2 + \left(4-\frac{13}{3}\right)^2$$

$$= \underline{\frac{844}{15}}$$

c) Iteration 1 :

ID	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
cluster centroid	C_3	C_1	C_1	C_1	C_1	C_2	C_2	C_2

$$\text{new } C_1 = \left(\frac{2+2+4+5}{4}, \frac{5+10+9+8}{4}\right) = \left(\frac{13}{4}, 8\right)$$

$$\text{new } C_2 = \left(7, \frac{13}{3}\right)$$

$$\text{new } C_3 = (1, 2)$$

Iteration 2 :

ID	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
cluster centroid	C_3	C_3	C_1	C_1	C_1	C_2	C_2	C_2

$$\text{new } C_1 = \left(\frac{2+4+5}{3}, \frac{10+9+8}{3}\right) = \left(\frac{11}{3}, 9\right)$$

$$\text{new } C_2 = \left(7, \frac{13}{3}\right)$$

$$\text{new } C_3 = \left(\frac{1+2}{2}, \frac{2+5}{2}\right) = (1.5, 3.5)$$

Iteration 3 :

ID	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
cluster centroid	C_3	C_3	C_1	C_1	C_1	C_2	C_2	C_2

cluster membership does not change => same centroids

final cluster centroids => $C_1 = \left(\frac{11}{3}, 9\right)$, $C_2 = \left(7, \frac{13}{3}\right)$, $C_3 = \left(\frac{3}{2}, \frac{7}{2}\right)$

$$SSE = \sum_{i=1}^{3} \sum_{x \in C_i} dist^2(c_i, x)$$

$$= \left(1-\frac{3}{2}\right)^2 + \left(2-\frac{7}{2}\right)^2 + \left(2-\frac{3}{2}\right)^2 + \left(5-\frac{7}{2}\right)^2 + \left(2-\frac{11}{3}\right)^2 + (10-9)^2$$

$$+ \left(4-\frac{11}{3}\right)^2 + (9-9)^2 + \left(5-\frac{11}{3}\right)^2 + (8-9)^2 + (6-7)^2 + \left(4-\frac{13}{3}\right)^2$$

$$+ (7-7)^2 + \left(5-\frac{13}{3}\right)^2 + (8-7)^2 + \left(4-\frac{13}{3}\right)^2$$

$$= 12$$

much lower than $k=2$

CL2

permutation of assignment of each centroid to cl...

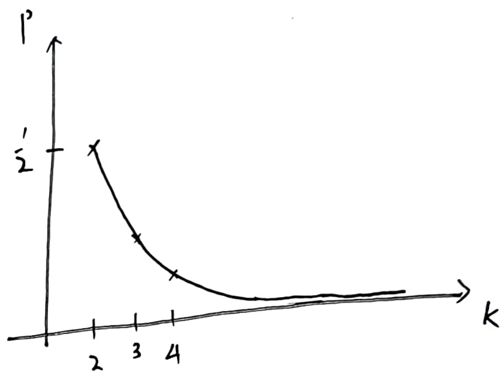
$$p = \frac{\#\text{ways to select 1 centroid from each cluster}}{\#\text{ways to select } k \text{ centroids}} = \frac{k!\, n^k}{(kn)^k} = \frac{k!}{k^k}$$

a) By Stirling's approximation

$$k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

$$\frac{k!}{k^k} \sim \frac{\sqrt{2\pi k}}{e^k}$$

$$\lim_{k \to \infty} \frac{k!}{k^k} = 0$$



b)

$\#$ways to select $2k$ centroids $= (kn)^{2k}$

$\#$ways to select at least 1 centroid from each cluster $=$

$$p \approx \left(\frac{k-1}{k}\right)^{2k}$$

$k=10$

$k=100$

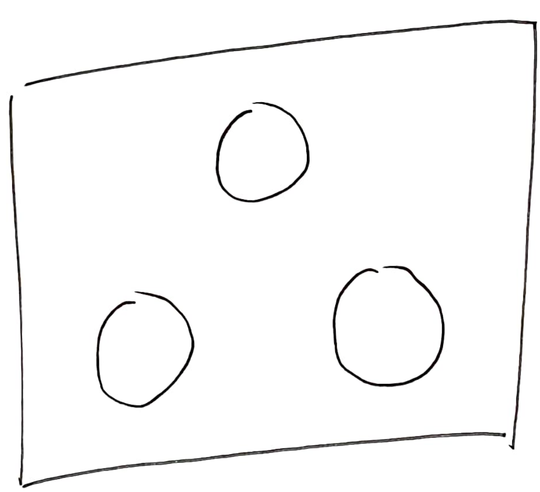
$k=1000$

CL3

The set of k points in the Voronoi diagram is similar to the k centroids in k-means cluster.
k clusters are formed by assigning each point to the closest centroid.
The regions/partitions in the voronoi diagram are the bounds for each cluster.
In k-means, the centroids will be recomputed based on the points in each cluster
while the voronoi diagram has these k points fixed.

k means is stochastic while a voronoi diagram is deterministic

CL4



equidistant clusters
are harder for bisecting
k means to split into
original 3 clusters