ROUGH K-MEANS CLUSTERING ALGORITHM

Algorithm: Rough K-Means

Input: Dataset of n objects with d features, number of clusters K and values of parameters W_{lower}, W_{upper}, and Epsilon.

Output: Lower approximation $\underline{U}(K)$ and Upper approximation $\overline{U}(K)$ of K Clusters.

Step1: Randomly assign each data object one lower approximation $\underline{U}(K)$. By definition (property 2) the data object also belongs to upper approximation $\overline{U}K$ of the same Cluster.

Step 2: Compute Cluster Centroids C_j

If $\underline{U}(K) \neq \emptyset$ and $\overline{U}(K)$ - $\underline{U}(K) = \emptyset$

$$C_j = \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|}$$

Else If $\underline{U}(K) = \emptyset$ and $\overline{U}(K)$ - $\underline{U}(K) \neq \emptyset$

$$C_j = \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$

Else

$$C_j = W_{lower} \times \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|} + W_{upper} \times \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$

Step 3: Assign each object to the lower approximation $\underline{U}(K)$ or upper approximation $\overline{U}(K)$ of cluster i respectively. For each object vector x, let $d(X, C_j)$ be the distance between itself and the centroid of cluster C_j.

$$d(X, C_j) = min_{1 \leq j \leq K}\, d(X, C_j).$$

The ratio $d(X, C_i) / d(X, C_j)$, $1 \leq i, j \leq K$ is used to determine the membership of x as follow: If $d(X, C_i) / d(X, C_j) \leq$ epsilon, for any pair (i, j), the $x \in \overline{U}(C_i)$ and $x \in \overline{U}(C_j)$ and x will not be a part of any lower approximation. Otherwise, $x \in \underline{U}(C_i)$, such that $d(X, C_i)$ is the minimum of $1 \leq i \leq K$. In addition $x \in \overline{U}(C_i)$.

Step 4: Repeat Steps 2 and 3 until convergence.

Illustrative Example

Table 1 shows example information system with real-valued conditional attributes. It consists of six objects/genes, and two features/samples. k = 2, which is the number of clusters. Weight of the lower approximation W_{lower}=0.7, Weight of the upper approximation W_{upper} = 0.3 and Relative threshold = 2.

Table 1 Example dataset for Rough K-Means

U	X	Y
1	0	3
2	1	3
3	3	1
4	3	0.5
5	5	0
6	6	0

Step1: Randomly assign each data objects to exactly one lower approximation

$\underline{K}_1$ = {(0, 3), (1, 3), (3, 1)}

$\underline{K}_2$ = {(3, 0.5), (5, 0), (6, 0)}

Step 2: In this case $\underline{U}(K) \neq \emptyset$ and $\overline{U}(K)$ - $\underline{U}(K) = \emptyset$, so we compute the centroid using $C_j = \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|}$,

$$C_1 = \left(\frac{0+1+3}{3}, \frac{3+3+1}{3}\right) = (1.33, 2.33)$$

$$C_2 = \left(\frac{3+5+6}{3}, \frac{0.5+0+0}{3}\right) = (4.67, 0.17)$$

Find the distance from centroid to each point using equlidean distance,

$$D_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$d_1(X, C_1)$:

(0, 3)(1.33, 2.33) $\Rightarrow \sqrt{(1.33-0)^2+(2.33-3)^2} = 1.49$

(1, 3)(1.33, 2.33) $\Rightarrow \sqrt{(1.33-1)^2+(2.33-3)^2} = 0.75$

(3, 1)(1.33, 2.33) $\Rightarrow \sqrt{(1.33-3)^2+(2.33-1)^2} = 2.13$

(3, 0.5)(1.33, 2.33) $\Rightarrow \sqrt{(1.33-3)^2+(2.33-0.5)^2} = 2.48$

(5, 0)(1.33, 2.33) $\Rightarrow \sqrt{(1.33-5)^2+(2.33-0)^2} = 4.45$

(6, 0)(1.33, 2.33) $\Rightarrow \sqrt{(1.33-6)^2+(2.33-0)^2} = 5.22$

$d_2(X, C_2)$:

(0, 3)(4.67, 0.17) $\Rightarrow \sqrt{(4.67-0)^2+(0.17-3)^2} = 5.46$

(1, 3)(4.67, 0.17) $\Rightarrow \sqrt{(4.67-1)^2+(0.17-3)^2} = 4.63$

(3, 1)(4.67, 0.17) $\Rightarrow \sqrt{(4.67-3)^2+(0.17-1)^2} = 1.86$

(3, 0.5)(4.67, 0.17) $\Rightarrow \sqrt{(4.67-3)^2+(0.17-0.5)^2} = 1.70$

(5, 0)(4.67, 0.17) $\Rightarrow \sqrt{(4.67-5)^2+(0.17-0)^2} = 0.37$

(6, 0)(4.67, 0.17) $\Rightarrow \sqrt{(4.67-6)^2+(0.17-0)^2} = 1.34$

Step 3: Assign each object to the lower approximation $\underline{U}(K)$ or upper approximation $\overline{U}(K)$ of cluster *i* respectively. Check If $d(X, C_i) / d(X, C_j) \leq$ epsilon.

1. (0, 3) $\Rightarrow d_2 / d_1 = 5.46/1.49 = 3.66443 \nleq 2$. So, x_1 will be a part of $\underline{K_1}$
2. (1, 3) $\Rightarrow 4.63/0.75 = 6.173 \nleq 2$. So, x_2 will be a part of $\underline{K_1}$
3. (3, 1) $\Rightarrow 2.13/1.86 = 1.145 < 2$, so x_3 will not be a part of $\underline{K_1}$ & $\underline{K_2}$
4. (3, 0.5) $\Rightarrow 2.48/1.70 = 1.458 < 2$, so x_4 will not be a part of $\underline{K_1}$ & $\underline{K_2}$
5. (5, 0) $\Rightarrow 4.35/0.37 = 11.756 \nleq 2$. So, x_5 will be a part of $\underline{K_2}$
6. (6, 0) $\Rightarrow 5.22/1.34 = 3.895 \nleq 2$. So, x_6 will be a part of $\underline{K_2}$

Now, we have clusters

$\underline{K_1}$= {(0, 3), (1, 3)} $\overline{K}_1$ = {(0, 3), (1, 3), (3, 1), (3, 0.5)}

$\underline{K_2}$= {(5, 0), (6, 0)} $\overline{K}_2$ = {(5, 0), (6, 0), (3, 1), (3, 0.5)}

Here, $\underline{U}(K) \neq \emptyset$ and $\overline{U}(K)$ - $\underline{U}(K) \neq \emptyset$ then find out the new centroid by using below equation,

$$C_j = W_{lower} \times \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|} + W_{upper} \times \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$

$$C_1 = 0.7 \times \left(\frac{0+1}{2}, \frac{3+3}{2}\right) + 0.3 \times \left(\frac{3+3}{2}, \frac{1+0.5}{2}\right) = (1.25, 2.325)$$

$$C_2 = 0.7 \times \left(\frac{5+6}{2}, \frac{0+0}{2}\right) + 0.3 \times \left(\frac{3+3}{2}, \frac{1+0.5}{2}\right) = (4.75, 0.225)$$

Step 4: Repeat Steps 2 and 3 until convergence (Old Centroid = New Centroid).