MACHINE INTELLIGENCE

Dr. N MEHALA
Department of Computer Science
and Engineering

Agenda

- K-Means Clustering Algorithm : Example

K-Means Clustering Algorithm

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-Means Clustering: Example

- Suppose we have 4 types of medicines and each medicine/object have two attributes or features as shown in the table. Our goal is to group these objects into K=2 group of medicine based on the two features (pH and weight index).

Object	Attribute 1 (X): weight index	Attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

K-Means Clustering: Example

Step1. Initial value of centroids: Suppose we use medicine A and medicine B as the initial centroids. Let c1 and c2 **denote the coordinate of the centroids, then c1 = (1,1) and c2 = (2,1)**

Object	Attribute 1 (X): weight index	Attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

K-Means Clustering: Example

Step2a. ***Objects-Centroids distance:*** *we calculate the distance between cluster centroid to each object. Let us* use ***Euclidean distance***, then we have distance matrix at iteration 0 is

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix} \qquad D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

- *First row of the distance matrix corresponds to the distance of each object to the first centroid*
- *Second row is the distance of each object to the second centroid*

Distance from the MedicineC =(4,3) to the first centroid C_1=(1,1):

$$\sqrt{(4-1)^2+(3-1)^2} = 3.61$$

Distance from the MedicineC =(4,3) to second centroid C_2=(2,1):

$$\sqrt{(4-2)^2+(3-1)^2} = 2.83$$

K-Means Clustering: Example

Step2b. ***Objects clustering:*** *we assign each object based on minimum distance. Thus medicineA is assigned to group1, medicineB is assigned to group2, medicineC is assigned to group2 and medicineD is assigned to group2. The element of Group matrix is 1 if and only if the object is assigned to that group.*

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

K-Means Clustering: Example

Step3. ***Recompute Centroids:*** Knowing the members of each group, now we compute the new centroid of each group based on these new memberships. Group1 only has one member thus the centroid remains in C1=(1,1). Group 2 now has three members, thus the centroid is the average coordinate among the three members: C2=$\left(\left(\frac{2+4+5}{3}\right),\left(\frac{1+3+4}{3}\right)\right)=\left(\frac{11}{3}\right),\left(\frac{8}{3}\right)$

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

MACHINE INTELLIGENCE

K-Means Clustering: Example (Iteration1)

PES UNIVERSITY ONLINE

K-Means Clustering: Example (Iteration1

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

Step2a. ***Objects-Centroids distance***

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

Step2b. ***Objects clustering***

$$\mathbf{G}^1 = \begin{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{matrix} group-1 \\ group-2 \end{matrix} \\ A \quad B \quad C \quad D & \end{matrix}$$

Step3. ***Recompute Centroids***

$$\mathbf{c}_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\tfrac{1}{2}, 1) \text{ and } \mathbf{c}_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$

K-Means Clustering: Example (Iteration2)

Step2a. ***Objects-Centroids distance***

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) \quad group-1 \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) \quad group-2 \end{matrix}$$

Step2b. ***Objects clustering***

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$
$$A \quad B \quad C \quad D$$

$$\mathbf{G}^2 = \mathbf{G}^1$$

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

K-Means Clustering: Evaluation

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

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

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

K-Means Clustering: Summary

- Pros: Easy to implement
- Cons: Can converge at local minima; slow on very large datasets
- Works with: Numeric values

K-Means Clustering: Summary

General approach to k-means clustering

1. Collect: Any method.
2. Prepare: Numeric values are needed for a distance calculation, and nominal values can be mapped into binary values for distance calculations.
3. Analyze: Any method.
4. Train: Doesn't apply to unsupervised learning.
5. Test: Apply the clustering algorithm and inspect the results. Quantitative error measurements such as sum of squared error can be used.
6. Use: Anything you wish. Often, the clusters centers can be treated as representative data of the whole cluster to make decisions

K-Means Clustering: Summary

- Initial centroids are often chosen randomly
  - Clusters produced vary from one run to another
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation etc.
- Most of the convergence happens in the first few iterations
  - Often the stopping condition is changed to 'Until relatively few points change clusters'
- Let, n = number of points, K = number of clusters, I = number of iterations, d = number of attributes
  - Complexity is O(n * K * I * d)

Conclusions

- K-Means Clustering (Partitional Clustering)
- Example: K-Means Clustering
- K-Means Clustering: Evaluation

MACHINE INTELLIGENCE

Resources

- http://www2.ift.ulaval.ca/~chaib/IFT-4102-7025/public_html/Fichiers/Machine_Learning_in_Action.pdf
- http://wwwusers.cs.umn.edu/~kumar/dmbook/.
- ftp://ftp.aw.com/cseng/authors/tan
- http://web.ccsu.edu/datamining/resources.html

THANK YOU

Dr. N MEHALA

Department of Computer Science and Engineering

mehala@pes.edu