UNIT - 3

Prof. Ranjitha U.N

CLUSTERING: Introduction

Clustering is the process of grouping the data into classes or clusters, so that the objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.

Although classification is an effective means for distinguishing groups, it requires costly collection and labelling of large data set.

Clustering is an example of unsupervised learning, which do not depend on predefined classes and class labels.

Hence clustering is a form of learning by observation rather than learning by examples.

K-Means clustering :-

K Means is a clustering algorithm for findings groups in the data,

the groups of data (cluster) are represented by their centers.

For the sample $X = \{x^t\}_{t=1}^{N}$

Consider m_i ~~$= x^t$~~ is an approximated value of x^t

The error in approximation should be as minimum as possible.

The Error $E(\{m_i\}_{i=1}^{k} | X) = \sum_t \sum_i b_i^t \| x^t - m_i \|^2$

where $b_i^t = \begin{cases} 1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\ 0 & \text{otherwise} \end{cases}$

* K-means algorithm iteratively calculate b_i^t for all x^t.
  $b_i^t = 1 \Rightarrow x^t$ belongs to group m_i
* with new x^t being added to m_i, b_i^t changes and needs to be recalculated.
* these steps are repeated until m_i stabilizes

K-means algorithm

1. Initialize m_i, $i = 1, \ldots k$ for example, to k random x^t.
2. Repeat
   For all $x^t \in X$
   $b_i^t \begin{cases} 1 & \text{if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\ 0 & \text{otherwise} \end{cases}$
   For all m_i, $i = 1 \ldots k$
   $m_i \leftarrow \dfrac{\sum_t b_i^t x^t}{\sum_t b_i^t}$
   Until m_i converge

Prof. Pojitha V N

Disadvantage : K-Means algorithm is a local search procedure and the final m_i highly depend on the initial m_i.

Methods to overcome this disadvantage

→ Considering randomly selected k instances as initial m_i

→ Calculating principal component, dividing its range, partitioning the data & take means of groups as initial centres

etc.

Note: Best way is to initialize centres where there is data.

PROBLEM

* Given the dataset of medicines, group it to relevant medicine cluster. Apply k-means for K = 2.

Solu:

Medicine	attrib1	attrib2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

consider any two data as the first centroids.

Let us consider

medicine A = first cluster, C_1 ⇒ group1

medicine C = Second cluster, C_2 ⇒ group2

step1 Considering cluster/group1 of medicine A ie $C_1 = (1, 1)$

calculate Euclidean distance.

A → $\sqrt{(1-1)^2 + (1-1)^2} = 0$

B → $\sqrt{(1-1)^2 + (0-1)^2} = 1$

C → $\sqrt{(0-1)^2 + (2-1)^2} = 1.41$

D → $\sqrt{(2-1)^2 + (4-1)^2} = 3.2$

E → $\sqrt{(3-1)^2 + (5-1)^2} = 4.5$

III^rd for $c_2 \cdot (0,2)$

A → $\sqrt{(1-0)^2 + (2-1)^2} = 1.4$

B → $\sqrt{(1-0)^2 + (2-0)^2} = 2.2$

C → $\sqrt{(0-0)^2 + (2-1)^2} = 0$

D → $\sqrt{(2-0)^2 + (4-2)^2} = 2.8$

E → $\sqrt{(3-0)^2 + (5-2)^2} = 4.2$

The new table

	For c_1	For c_2		
A	0	1.4	→ belongs to	c_1
B	1	2.2	→ " "	"
C	1.41	0	→ " "	c_2
D	3.2	2.8	→ "	
E	4.5	4.2	→ "	

Due to new grouping, the centroid also changes

New $c_1 = \frac{1+1}{2}, \frac{1+0}{2}$ New $c_2 = \frac{0+2+3}{3}, \frac{2+4+5}{3}$

$c_1' = (1, 0.5)$ $c_2' = (1.7, 3.7)$

<u>Step 2</u> Consider c_1' & c_2' and compute Euclidean distance for given data.

we get

	For c_1	For c_2		
A	0.5	2.7	→ belongs to	c_1
B	0.5	3.7	→ "	c_1
C	1.8	2.4	→ "	c_1
D	3.6	0.5	→ "	c_2
E	4.9	1.9	→ "	c_2

New centroid

$c_1^2 = \frac{1+1+0}{3}, \frac{1+0+2}{3}$ $c_2^2 = \frac{2+3}{2}, \frac{4+5}{2}$

$= (0.7, 1)$ $= (2.5, 4.5)$

) Apply KNN Means clustering on the given dataset for k = 2.

K = 2 , D = {2, 3, 4, 10, 11, 12, 20, 25, 30}

Soln Taking two means randomly (if not given)

Initialize $m_1 = 4$ $m_2 = 12$

$D_1 = \{2, 3, 4\}$ $D_2 = \{10, 11, 12, 20, 25, 30\}$

Step 1: $m_1' = \frac{2+3+4}{3} = 3$ $m_2' = \frac{108}{6} = 18$

Step 2: consider the {D} for new m_1' & m_2'

$D_1 = \{2, 3, 4, 10\}$ $D_2 = \{11, 12, 20, 25, 30\}$

$m_1'' = 4.75$ $m_2'' = 19.6$

≈ 5 ≈ 20

Step 3: consider {D} for new m_1'' & m_2''

$D_1 = \{2, 3, 4, 10, 11, 12\}$ $D_2 = \{20, 25, 30\}$

$m_1''' = \frac{2+3+4+10+11+12}{6}$ & $m_2''' = \underline{25}$

$= \underline{7}$

Step 4: consider {D} for m_1''' & m_2'''

$D_1 = \{2, 3, 4, 10, 11, 12\}$ $D_2 = \{20, 25, 30\}$

$m^4 = \underline{7}$ $m^4 = \underline{25}$

Since we are getting same mean, the grouping or clustering changes no further.

Hence KNN concludes.

Step 3 : continuing the same process for new centroids c_1^2 & c_2^2

we get

A	0.3	3.8	c_1
B	1.04	4.7	c_1
C	1.22	3.5	c_1
D	3.3	0.7	c_2
E	4.6	0.7	c_2

The clustering/grouping does not change & hence the centroids also do not change.

Hence K-means converges

Ranjitha U.N

Supervised learning after Clustering

* Clustering can be used for two purposes.
  → Used for data exploration, to understand structure of the data.
  → Used to find similarities between instances and thus group instances.
* The mean of the cluster group formed gives the representative prototype of instances in the group.
  Ex: Consider a cluster which is formed for the sales of a product in a particular region.
  An instance from that group will tell the requirement of the people of the region (CRM).
* Clustering is also used as a preprocessing stage.
* The advantage of having a unsupervised ~~learning~~ Clustering before supervised learning is that unsupervised learning does not need labelled data and labelled data is expensive.

Hierarchial clustering

* Hierachial clustering aims at finding groups such that instances in a group are more similar to each other than instances in different groups.

* Hierarchical clustering makes use of the Euclidean distance measure.

Euclidean distance $d(a,b) = \sqrt{(d_a - d_b)^2}$, which is a special case of Minkowski distance with $P = 2$

$$d_m(x^r, x^s) = \left[\sum_{j=1}^{d} (x_j^r - x_j^s)^p \right]^{1/p}$$

other distance measure is city-block distance (Refer text)

* Hierachial clustering approach has various types viz agglomerative clustering algorithm, divisive clustering, single link clustering, complete link clustering.

→ Agglomerative clustering:

1. starts with N groups, each initially containing one training instance, then merging similar groups to form larger groups, until there is a single one.
2. At each iteration, two closest groups are chosen to merge.
3. The result of Agglomerative clustering is a hierarchial structure called dendrogram.
4. Dendrogram represent a tree, where leaves corresponds to instances which are grouped in the order of their merge.

→ Divising clustering algorithm:-

1. It starts with a large group, and deviding large groups into smaller groups, until each group contains a single instance.

→ single link clustering:

1. The distance between instance used for grouping is defined as the smallest distance between all possible pair of elements of the two groups.

$$d(G_i, G_j) = \min_{x^r \in G_i, x^s \in G_j} d(x^r, x^s)$$

2. If we consider a weighted, completely connected graph with nodes being instances and the edges with weights being distance b/w instances. Then single link method corresponds to constructing minimal spanning tree.

→ complete-link clustering

1. the distance between two groups is taken as the largest distance between all possible pairs.

$$d(G_i, G_j) = \max_{x^r \in G_i, x^s \in G_j} d(x^r, x^s)$$

single link cluster fig 7.5 from text.

Choosing the Number of clusters :-

* The complexity of clustering depends on the no. of clusters 'k'.

* There are various ways to fine tune k, such as.

1. In case of colour quantization, k is defined by the application.
2. Plotting the data in 2-D using PCA.
3. An incremental approach, where maximum allowed distance is selected and made equivalent to maximum allowed reconstruction error.
4. Manual check on clusters being meaningful groups of data.

* Reconstruction error can be plotted as a function of 'k' depending on the type of the clustering method used.

* In hierarchial clustering, the difference between levels in the tree is considered to decide on a good split.

Expectation - Maximization Algorithm. (EM)

Prof. Pavithra V

* EM works in a similar fashion to that of K-means, but EM yields a soft decision (elliptical curves) and k-means yields a hard decision (0/1, circles).

* K-means is applied on models of data which are independent.
  EM is applied on a mixture model.

Note: Mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without subpopulation identity information. ⇒ Latent variables. (z_i)

* Expectation-Maximization algorithm's approach is
  1. To find maximum likelihood of parameters in statistical model (where the model depends on unobserved latent variables ⇒ Mixture model). the expectation (E) step which creates log likelihood function
  2. The Maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E-step.

* EM algorithm is an iterative method which alternates between E & M.

* The mixture density of a mixture model is given by

$$P(x) = \sum_{i=1}^{K} p(x|G_i)\, P(G_i) \quad \text{---} (1)$$

eq $G_i \rightarrow$ mixture components

$p(x|G_i) \rightarrow$ component densities

Using (1), For the given sample $X = \{x^t\}_t$, the log likelihood is

$$\mathcal{L}(\Phi|X) = \log \prod_t p(x^t|\Phi)$$

$$= \sum_t \log \sum_{i=1}^{K} p(x^t|G_i)\, P(G_i) \quad \text{---} (2)$$

* the goal of EM algorithm is to find vector Φ that maximizes the likelihood of the observed values of X. Since z is also the parameter hidden in the model, likelihood $\mathcal{L}_c$ is

$$\mathcal{L}_c(\Phi|X, z) \quad \text{---} (3)$$

* Since z values are not observed, L_c, cannot be worked, hence working with its expectation Q,

$$\text{E step}: Q(\Phi|\Phi^l) = E\left[\mathcal{L}_c(\Phi|X, z)\,|\,X, \Phi^l\right] \quad \text{---} (4)$$

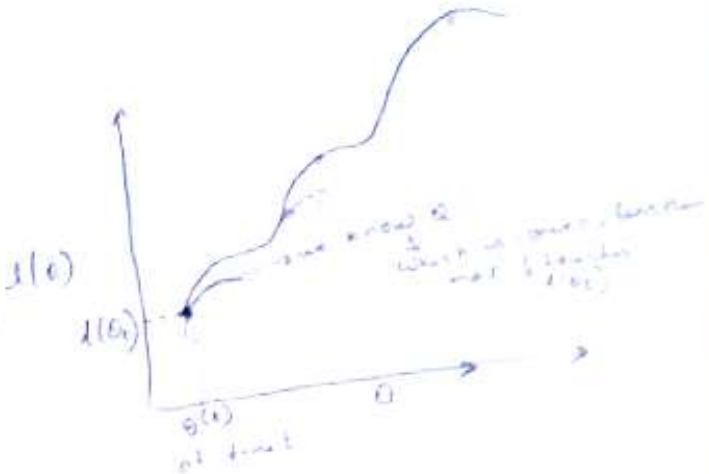
Φ^l = current parameter value, l indexes iteration

In maximization step (M), new parameter value ϕ^{t+1} is found, that maximizes Eq (4)

Thus,

$$M \text{ step}: \quad \phi^{t+1} = \arg\max_{\phi} Q(\phi|\phi^t) \quad \text{—(5)}$$

Illustration



θ to be found where $l(\theta)$ is max

approach is gradient based

then it is [illegible]

θ

θ^{t+1} : $l(\theta^t)$

$\theta^t(\theta) \leq l(\theta)$ ← bounded by our function

Finding θ → maximized

→ [illegible]

→ [illegible]

while not converged

[illegible]

[illegible]