Introduction to clustering

Using python.

Iván Andrés Trujilllo Abella

Aicoll

Unidad de analítica

Insights

How we can measure if two samples are similar? what are the primary focus on latex

new frame

a new frame will be important for us

Challenges

- How many groups we can find?
- How choose relevant variables?

Clustering

It is a optimization problem. That involves similarity among features. the most uses measure it is a distance metric among two points.

Data

Economy	PIB	Mean Growth
A	10	0.5
B	11	0.7
C	12	1.2
D	14	0.3

Table: Solow hypothesis

Euclidean distance

The distance as a approximation to similarity.

$$d_{ij} = \sqrt{\sum (x_{if} - x_{jf})^2} \tag{1}$$

where f indicate the feature of the individuals ij

	A	**B**	**C**	**D**
A	0	d_{AB}	d_{AC}	d_{AD}
B	d_{BA}	0	d_{BC}	d_{BD}
C	d_{AC}	d_{BA}	0	d_{CD}
D	d_{AD}	d_{BD}	d_{DC}	0

Table: Euclidean distance matrix

note the symmetry $d_{AB} = d_{BA} = \sqrt{(11-10)^2 + (0.7-0.5)^2}$.

Association coefficients

		B	
		Feature	**Not feature**
A	**Feature**	a	b
	Not feature	c	d

$S_{(ij)} = \frac{a+d}{a+b+c+d}$ take in mind that two objects could be similar by lacking feature the following could be tackle this problem $J_{(ij)} = \frac{a}{a+b+c}$. Notice that the both are numbers between zero and one, the first indicate not similarity a

Methods of clustering

Hierarchical clustering and k-means, are most popular methods to clustering.

Hierarchical cluster

```
n points then n cluster:
find the most pair similar cluster and merge
(step by step namely will be one fewer):
stop when all points are merged in one cluster
```

Linkage

if we have more of one point how measure?

- single: the shortest distance between two any member of two clusters.

$$d(C_i, C_j) = min\{d(i,j)\}, \forall i,j \in C_i x C_j \tag{2}$$

- Complete: the greatest distance from any member to another member.

$$d(C_i, C_j) = max\{d(i,j)\}, \forall i,j \in C_i x C_j \tag{3}$$

- Avarage: Consider the mean of distances among the points of clusters.

$$d(C_i, C_j) = d(C_{\bar{x_i}}, C_{\bar{x_j}}). \tag{4}$$

Stopping criteria

- Minimun number of clusters: reach a minimum number of cluster
- treshold of maximun distance: not joint cluster with a maximun distance
- maximun of steps:

k means

We can make a partition of n individuals in k groups, and denote $p(n,k)$ the distance of the point i to the c

$$d_{i,c} = (\sum_{f=1}^{m}(x_{i,f} - \bar{x}_{c,f}) \tag{5}$$

therefore:

$$e(p(n,k)) = \sum d_{i,c}^2 \tag{6}$$

Now we must select the arrangement that minimize $e(p(n,k)$.

K means

```
chose k initial centroids:
 assing each observation to the closest centroid
 assing new centroids
 break the assingantion if not change
```

How update the centroids

Suppose that you consider N variables, and k cluster therefore,

$$C_i = (\bar{x}_{1i}, \bar{x}_{2i}, ...\bar{x}_{Ni}), i = 1, 2, ..., k \tag{7}$$

Remember that i denote the cluster actually assigned then the calculate is over all points that belong to the cluster $\forall j \in S_i$. This process remain until not change the composition of clusters.

Complexity

k cluster for each p points and t time of calculate the metric.

Problems

Sensible to the selection of k.

Question

the result depend upon initial centroids?

Choose k

θ observations in k groups, $2 < k < \theta$

- A prior knowledge
- Iteration
- Uses hierarchical cluster

The reduction of the number of cluster imply lost in homogeneity.

Choose k

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} W_{(i,j)} \|X^i - \mu^j\|_2^2 \tag{8}$$

remember that $\boldsymbol{x}$ and $\boldsymbol{y}$

Inertia

$$
\begin{aligned}
SSE &= \sum_{i}^{n}\sum_{j}^{k} W_{(i,j)} d(x_i, c_j)^2 \\
&= \sum_{i}^{n} [W_{i,1} d(x_i, c_1)^2 + W_{i,2} d(x_i, c_2)^2 + ... + W_{(i,k)} d(x_i, c_k)] \\
&= \sum_{i}^{n} W_{(i,1)} d(x_i, c_1)^2 + \sum_{i}^{n} W_{(i,2)} d(x_i, c_2) + ... + \sum_{i}^{n} W_{(i,k)} d(x_i, c_k) \\
&= W_{(1,1)} d(x_1, c_1) + ... + W_{(n,1)} d(x_n, c_1) + \\
&\quad W_{(1,2)} d(x_1, c_2) + ... + W_{(n,2)} d(x_n, c_2) + \\
&\quad W_{(1,k)} d(x_1, c_k) + ... + W_{(n,k)} d(x_n, c_k)
\end{aligned}
\tag{9}
$$

Assessment of quality

Silhouette is a measure that give us a number from -1 to 1.

$$s^i = \frac{b^i - a^i}{\max(b^i, a^i)} \tag{10}$$

a^i the average distance among a sample that $x \in i$ and the other samples of the same group.
b^i the average distance among $x \in i$ and the all other samples of the closest group.
how values of s^i are ideal?,

Fuzzy c means clustering

Each point have a membership value to each cluster.

$$\sum_{k=1}^{m}\sum_{j=1}^{n} f_{jk}^{2}\|x_j - \mu_k\| \tag{11}$$

take in mind that f_{jk} it is the is the membership value of the j individual in the k cluster.
u_k it is a function also of the points of data and membership values.

Cluster ideas

hard clustering: problems with no overlapping. soft clustering: belong to more than one centroid (K-means).
minimiza intra-clusters maximizing inter-cluster.

Examples of c fuzzy means

Cancer data analysis
Impact on industry
Segmentation cancer tissue

Until now

- spherical shapes with k-means
- stopping criteria with hierarchical

DBScan

We can trait noise with DBScan.
Works differently to another two:

- Density

Core object (r, η)

object that have at least η neighborhoods in a radius of r. think that a core object it is a candidate point to be a cluster.

H object

we said that a pattern or point H is **directly reachable** from a another point O if H it is neighbor of O and O it is object core.

S object

We said that a pattern or point S is **indirectly reachable** from another point O if there are a sequence of of objects $p_1, p_2, ..., p_n$ where p_i is directly reachable from p_{i-1}. where $p_1 = O$ and $p_n = S$.
To chain is apply to core objects.

summary in object core, border object and noise object.

Outliers

Outliers tend to have less densities.

Advantages

- we dont need provided the number of cluster as in K-means
- not is contingent to spherical shapes
- handled noise and outliers

Disadvantages

- rely on in the knowledge domain to tune the hyperparameters.