MAC 0459 / 5865

Data Science and Engineering

R. Hirata Jr. (hirata@ime.usp.br)

Class 12 (2020)

Errata para um dos slides da aula anterior

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} |\mathrm{x}_{ik} - \mathrm{x}_{jk}|^p\Big]^{1/p}$$

Errata para um dos slides da aula anterior

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} |\mathrm{x}_{ik} - \mathrm{x}_{jk}|^p \Big]^{1/p}$$

$p \geq 1$ **distance**

Errata para um dos slides da aula anterior

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} |\mathrm{x}_{ik} - \mathrm{x}_{jk}|^p \Big]^{1/p}$$

$p \geq 1$ **distance**

$0 < p < 1$ **not a distance but a similarity**

Clustering

Unsupervised learning - EDA technique

Given a set of objects with no associated labels, is there any pattern that can appear in the set? What can we learn/deduce from the pattern?

Data mining

Data mining is a buzzword

It is part of the task of **knowledge discovery** in large **databases** (KDD)

Data mining

Data mining is a buzzword

It is part of the task of **knowledge discovery** in large **databases** (KDD)

Usual tasks in KDD:

- detection of anormalities (changes, deviations)
- detection of association rules (dependencies between variables, structures, etc)
- **cluster analysis**
- classification
- summarization (compact representation, visualization, report)

Cluster analysis

- **Clustering** or **cluster analysis** is the main approach to unsupervised learning and a very important EDA approach.
- No information available but the dataset
- **Objective**: find groups or natural structures hidden in the dataset.
- The concept of **group** is vaguely defined; usually depends on another concept as, or instance, **similarity**.

Cluster analysis

- *Clustering* is usually associated with the process of computing or finding clusters in a dataset.

- Clustering or **cluster**

- A cluster is **good** if for any two objects in a cluster/group they are **more similar to each other than to any other object in the dataset that does not belong to that group.**

Clustering – definition

Notation:

$X = \{x_1, x_2, \ldots, x_m\}$ (objects)
Number of clusters: c (usually unknown)
Clusters: $C_1,\ C_2,\ \ldots,\ C_c$

Standard definition

Partition

A **partition of a set** X is a collection of parts/subsets $C_1, C_2, \ldots, C_c$, $c > 0$, such that:

- $C_j \neq \emptyset, \qquad j = 1, \ldots, c$
- $\cup_{j=1}^{c} C_j = X$
- $C_i \cap C_j = \emptyset, \qquad i, j = 1, 2, \ldots, c$ e $i \neq j$

The parts/subsets C_1, C_2, ..., C_c are know as

clusters,

groups,

parts

Fuzzy clustering

Fuzzy clustering: **objects can be part of one or more parts**;

- **membership degree** of an object to a cluster is given by a *membership function* $u_j : X \to [0,1], \quad j = 1, 2, \ldots, c,$

- for each object x_i, the sum of its membership degrees to all clusters is equal to 1:

$$\sum_{j=1}^{c} u_j(\mathrm{x}_i) = 1, \qquad i = 1, 2, \ldots, m$$

- for each cluster/part j, there exists an object whose membership degree is non zero

$$0 < \sum_{i=1}^{m} u_j(\mathrm{x}_i) < m, \qquad j = 1, \ldots, c$$

Hard ou crispy clustering

clustering is usually called **hard or crispy clustering**:

because an object is part of one and only one part/cluster,

for each $x \in X$, *there exists a* j *such that* $u_j(x) = 1$

Clustering approaches

- **Sequential:** fast and straightforward because the objects (feature vectors) are presented at most six times to the algorithm. The final result is dependent of the order of the objects presented. The resulting clusters are compact and hyperspherical or hyperellipsoidal.
- **Hierarchical:** agglomerative or divisive.
  - agglomerative: decreasing sequence of the number of clusters.
  - divisive: increasing sequence of the number of clusters.
- **Optimization:** The number is usually fixed and a cost function is optimized.
- **Other:** Branch and bound, genetic, stochastic, etc.

Clustering – workflow

- **Choose a representation for the objects (vector of characteristics, matrix, etc)**
- **Choose a clustering approach (hierarchical, iteractive, etc)**
- **Choose a measure to distinguish the objects**
- **Choose a measure to evaluate the clusters**
- **Choose an algorithm of clustering**
- **Validation**
- **Interpretation**

Applications

Applications

- **Data reduction**:
- **Hypothesis generation**:
- **Hypothesis testing** :
- **Prediction/classification based on groups**:

Applications

- **Data reduction**: depending of the amount of data, clusters can be used as representers;
- **Hypothesis generation**: infer some hypothesis concerning the nature of the data. The hypothesis must be verified using other datasets.
- **Hypothesis testing** : verify the validity of a specific hypothesis.
- **Prediction/classification based on groups**: given a clustering result and a new object, which cluster best represent this new object?

Distance, similarity, dissimilarity

- a positive number that represents the separations between two objects
  - S a set of objects
  - $D : S \times S \rightarrow \mathbb{R}^+$
- A distance function satisfies four properties: positivity, identity, simmetry and triangle inequality.
- A similarity does not need to satisfy the triangle inequality

Distance between objects, objects and groups, groups and groups

Notation:

$d(\mathrm{x}_i, \mathrm{x}_j)$: between objects x_i and x_j

$d(\mathrm{x}, C)$: between object x and group C

$d(C_i, C_j)$: between groups C_i and C_j

Distance between object and groups

Distance between an object x and a group of objects C:

- **Closest distance**

$$d(x, C) = \min_{y \in C} d(x, y)$$

- **Farest distance**

$$d(x, C) = \max_{y \in C} d(x, y)$$

- **Mean distance**

$$d(x, C) = \frac{1}{|C|} \sum_{y \in C} d(x, y)$$

Distance between object and groups

Find a representer to a group C, compute the distance between the representer and an object x:

Possible representer:

- **point** (make sense when the groups are spherical)
  - **mean vector (or point)**: $\mathrm{m}_p = \dfrac{1}{|C|}\sum_{\mathrm{y} \in C} \mathrm{y}$
  - **central point**: $\mathrm{m}_c \in C$ tal que $\sum_{\mathrm{y} \in C} d(\mathrm{y}, \mathrm{m}_c) \leq \sum_{\mathrm{y} \in C} d(\mathrm{y}, \mathrm{m}), \quad \forall \mathrm{m} \in C$
  - **median point** : m_{med} tal que $med\{d(\mathrm{y}, \mathrm{m}_{med}), \ \ \mathrm{y} \in C\} \leq med\{d(\mathrm{y}, \mathrm{m}), \ \ \mathrm{y} \in C\}, \ \ \forall \mathrm{m} \in C$
- **hyperplane, hypercurve** : the distance between an object x and the representer (for a group C), when the **cluster is represented by a hyperplane, or hypercurve** can be the distance between a point and the curve

Distance between groups

- **Single linkage**

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

- **Complete linkage**

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

- **Average linkage**

$$d(C_i, C_j) = \frac{1}{|C_i|\,|C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

Hierarchical clustering

Aglomerative hierarchical clustering

Algorithm:

Initial partition: $\mathcal{C}_0 = \{C_i = \{\mathrm{x}_i\}, \quad i = 1, 2, \ldots, m\}$

$t = 0$

Repete

$t = t + 1$

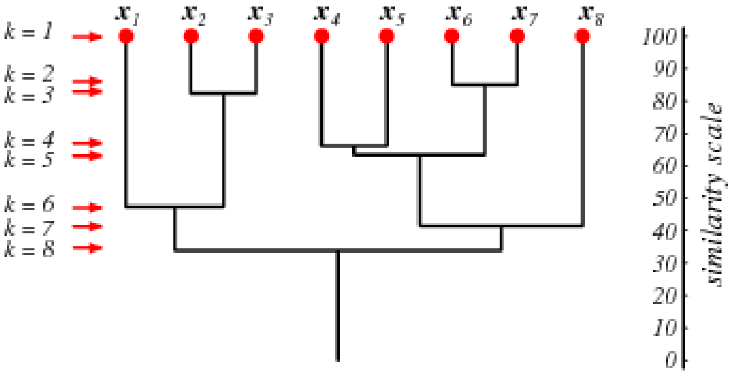
Among all possible pairs of clusters $\mathcal{C}_{t-1}$, find the pair, say C_i, C_j, with the best similarity.

$C_q = C_i \cup C_j$

$\mathcal{C}_t = (\mathcal{C}_{t-1} \setminus \{C_i, C_j\}) \cup \{C_q\}$

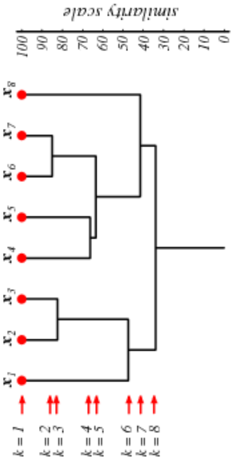
until $\mathcal{C}_t$ contains only one cluster.

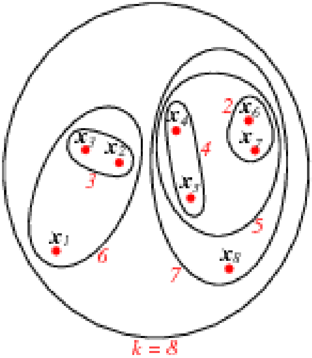
Aglomerative hierarchical clustering - dendrogram



k: iteraction step

Aglomerative hierarchical clustering - partition grouping





k: iteraction step

Aglomerative hierarchical clustering

Characteristics

Aglomerative hierarchical clustering

Characteristics

- **easy to understand**

Aglomerative hierarchical clustering

Characteristics

- **easy to understand**
- **memory intensive**

Aglomerative hierarchical clustering

Characteristics

- **easy to understand**
- **memory intensive**
- **can not undo the last step**

Aglomerative hierarchical clustering

Characteristics

- **easy to understand**
- **memory intensive**
- **can not undo the last step**
- **sensible to noise**

Aglomerative hierarchical clustering

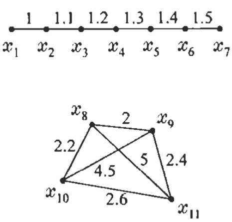
Characteristics

- **easy to understand**
- **memory intensive**
- **can not undo the last step**
- **sensible to noise**
- **can use other grouping criteria besides the ones based on linkage**

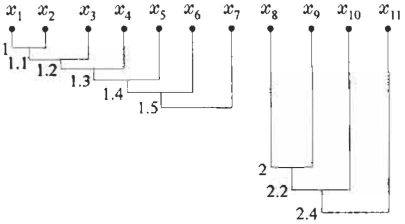
Aglomerative hierarchical clustering

Dendrograms when we vary the linkage criteria:
(a) dataset (distance between pairs)
(b) single linkage
(c) complete linkage

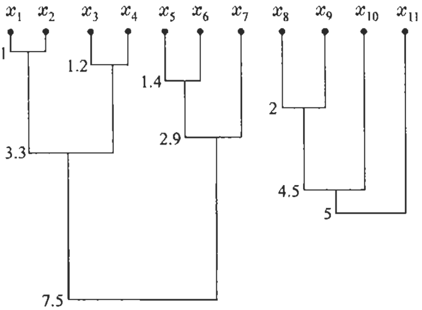
Check the next slide


1 1.1 1.2 1.3 1.4 1.5
x_1 x_2 x_3 x_4 x_5 x_6 x_7
x_8 2 x_9
2.2 5 2.4
4.5
x_{10} 2.6 x_{11}

(a)


x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11}
1
1.1
1.2
1.3
1.4
1.5
2
2.2
2.4

(b)


x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11}
1
1.2
1.4
2
2.9
3.3
4.5
5
7.5

(c)

Aglomerative hierarchical clustering

How to choose the number of clusters?

Aglomerative hierarchical clustering

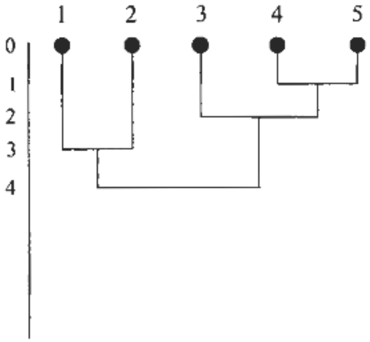
How to choose the number of clusters?

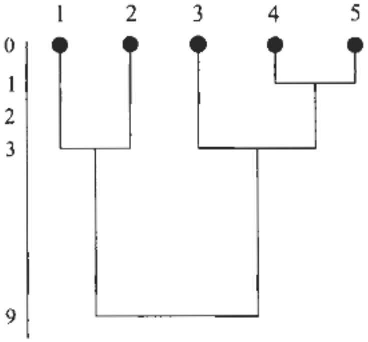
Considert the lifetime of a group, ie, the absolute value of the difference between two levels of similarity (when a group was created).

Aglomerative hierarchical clustering

How to choose the number of clusters?

Considert the lifetime of a group, ie, the absolute value of the difference between two levels of similarity (when a group was created).





Aglomerative hierarchical clustering

Lifetime analysis: subjective

Dendrogram visualization is impossible if the amount of data is large.

There are several different ways to evaluate a cluster quality, for instance the between-within classes.

Aglomerative hierarchical clustering

Single linkage: chaining effect.

Mean distance: clusters are more spherical and compact.

Computacional complexity ?