# Série de TD N°5

Clustering 2 (Clustering Hiérarchique)

## Exercice1 :

Le dendrogramme est une représentation graphique d’une classification (clustering) hiérarchique par un arbre.

1. Le dendrogramme d’une classification ascendante (ou descendante) est-il unique ? Si OUI dites Comment ? Si NON dites pourquoi ?
2. Comment déterminer le nombre de classes (clusters) à partir du dendrogramme ? Cette méthode est-elle exacte ou approximative ? Quel est le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité?
3. Considérer la matrice de similarité suivante de cinq documents d1, d2, d3, d4 et d5. Déterminer le dendrogramme résultant de l’application du text clustering hiérarchique ascendant en utilisant le « **lien maximum** ».

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **d1** | **d2** | **d3** | **d4** | **d5** |
| **d1** | 0 | 0.5 | 0.5 | 0.6 | 0.8 |
| **d2** | 0.5 | 0 | 0.7 | 0.6 | 0.5 |
| **d3** | 0.5 | 0.7 | 0 | 0.6 | 0.5 |
| **d4** | 0.6 | 0.6 | 0.6 | 0 | 0.9 |
| **d5** | 0.8 | 0.5 | 0.5 | 0.9 | 0 |

## Exercice2 :

Considérer la matrice de dissimilarité suivante P.

Déterminer les dendrogrammes résultants de l’application du

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **a** | **b** | **c** | **d** | **e** |
| 0 | 4 | 9 | 6 | 5 |
| 4 | 0 | 3 | 8 | 7 |
| 9 | 3 | 0 | 3 | 2 |
| 6 | 8 | 3 | 0 | 1 |
| 5 | 7 | 2 | 1 | 0 |

« **single link algorithm** », puis du « **complete link algorithm** » sur P et commentez.

**Exercice3 :**

|  |  |
| --- | --- |
| Considérons la matrice de similarité suivante entre les exemples  x1, x2, x3, x4 et x5  Déterminer le dendrogramme résultant de l’application du « single link algorithm » (lien minimum). | X1 0 0.4 0.9 0.6 0.5  X2 l0.4 0 0.3 0.8 0.7  X3 I0.9 0.3 0 0.3 0.2I  X4 0.6 0.8 0.3 0 0.1  X5 𝗁0.5 0.7 0.2 0.1 0 )  x1 x2 x3 x4 x5 |

# Série de TD N°5 (Corrigé) Clustering 2 (Clustering Hiérarchique)

## Exercice1 :

1. Le dendrogramme d’un clustering n’est pas unique ; il dépend de la stratégie de regroupement : lien minimum, maximum ou moyen. En plus, si la distance minimale choisie n’est pas unique, le choix aléatoire diversifie le dendrogramme.
2. On détermine le nombre de clusters à partir d’un dendrogramme, en trouvant le nombre de points d’intersection entre la droite *y=d* et le dendrogramme, *d* étant la distance choisie. Cette méthode est exacte. Le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité se fait généralement à l’aide de l’avis d’un expert dans le domaine.
3. Le dendrogramme résultant de l’application du text clustering hiérarchique ascendant de cinq documents d1, d2, d3, d4 et d5, en utilisant le « **lien maximum** ».

0.9

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **d1** | **d2** | **d3** | **d4** | **d5** |
| **d1** | 0 |  |  |  |  |
| **d2** | **0.5** | 0 |  |  |  |
| **d3** | 0.5 | 0.7 | 0 |  |  |
| **d4** | 0.6 | 0.6 | 0.6 | 0 |  |
| **d5** | 0.8 | 0.5 | 0.5 | 0.9 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **d1d2** | **d3** | **d4** | **d5** |
| **d1d2** | 0 |  |  |  |
| **d3** | 0.7 | 0 |  |  |
| **d4** | 0.6 | 0.6 | 0 |  |
| **d5** | 0.8 | **0.5** | 0.9 | 0 |

0.6

0.5

grouper (d1, d2) grouper (d3, d5)

0 d1

d2 d4

d3 d5

**Dendrogramme**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **d1d2** | **d3d5** | **d4** |
| **d1d2** | 0 |  |  |
| **d3d5** | 0.8 | 0 |  |
| **d4** | **0.6** | 0.9 | 0 |

|  |  |  |
| --- | --- | --- |
|  | **d1d2d4** | **d3d5** |
| **d1d2d4** | 0 |  |
| **d3d5** | **0.9** | 0 |

grouper (d1d2, d4) grouper(d1d2d4,d3d5)

## Exercice3 :

**Single Link (Lien mimimum)**

grouper (x4, x5)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x1** | **x2** | **x3** | **x4** | **x5** |
| **x1** | 0 |  |  |  |  |
| **x2** | 0.4 | 0 |  |  |  |
| **x3** | 0.9 | 0.3 | 0 |  |  |
| **x4** | 0.6 | 0.8 | 0.3 | 0 |  |
| **x5** | 0.5 | 0.7 | 0.2 | **0.1** | 0 |

grouper(x3,x4x5)

0.4

0.3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **x1** | **x2** | **x3** | **x4x5** |
| **x1** | 0 |  |  |  |
| **x2** | 0.4 | 0 |  |  |
| **x3** | 0.9 | 0.3 | 0 |  |
| **x4x5** | 0.5 | 0.7 | **0.2** | 0 |

0.2

0.1

0

x1 x2

x3 x4 x5

**Dendrogramme**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **x1** | **x2** | **x3x4x5** |
| **x1** | 0 |  |  |
| **x2** | 0.4 | 0 |  |
| **x3x4x5** | 0.5 | **0.3** | 0 |

|  |  |  |
| --- | --- | --- |
|  | **x1** | **x2x3x4x5** |
| **x1** | 0 |  |
| **x2x3x4x5** | **0.4** | 0 |

grouper (x2, x3x4x5)