S. Canu
Ph. Leray
A. Rakotomamonjy

Data-Mining

Corrigé Examen 2002/2003

4^{eme} année

1 Clustering (13 points)

X	1	2	9	12	20

1. (7 points) K-Means
   (a) Appliquez l'algorithme des K-means avec les valeurs de k et les points de départ suivants :
      i. $k = 2, \mu_1 = 1, \mu_2 = 20$.

	1	2	9	12	20
$d^2(x, \mu_1)$	0	1	64	121	361
$d^2(x, \mu_2)$	361	324	121	64	0

$\mu_1 = (1 + 2 + 9)/3 = 4; \mu_2 = (12 + 20)/2 = 16$

	1	2	9	12	20
$d^2(x, \mu_1)$	9	4	25	64	256
$d^2(x, \mu_2)$	225	196	49	16	16

μ_1 et μ_2 ne changent pas ==> convergence

      ii. $k = 3, \mu_1 = 1, \mu_2 = 12, \mu_3 = 20$.

	1	2	9	12	20
$d^2(x, \mu_1)$	0	1	64	121	361
$d^2(x, \mu_2)$	121	100	9	0	64
$d^2(x, \mu_3)$	361	324	121	64	0

$\mu_1 = (1 + 2)/2 = 1.5; \mu_2 = (9 + 12)/2 = 10.5; \mu_3 = 20$

	1	2	9	12	20
$d^2(x, \mu_1)$	0.25	0.25	56.25	110.25	342.25
$d^2(x, \mu_2)$	90.25	72.25	2.25	2.25	90.25
$d^2(x, \mu_3)$	361	324	121	64	0

μ_1, μ_2 et μ_3 ne changent pas ==> convergence

      iii. $k = 4, \mu_1 = 1, \mu_2 = 9, \mu_3 = 12, \mu_4 = 20$.

	1	2	9	12	20
$d^2(x, \mu_1)$	0	1	64	121	361
$d^2(x, \mu_2)$	64	49	0	9	121
$d^2(x, \mu_3)$	121	100	9	0	64
$d^2(x, \mu_4)$	361	324	121	64	0

$\mu_1 = (1 + 2)/2 = 1.5; \mu_2 = 9; \mu_3 = 12; \mu_4 = 20$

	1	2	9	12	20
$d^2(x, \mu_1)$	0.25	0.25	56.25	110.25	342.25
$d^2(x, \mu_2)$	64	49	0	9	121
$d^2(x, \mu_3)$	121	100	9	0	64
$d^2(x, \mu_4)$	361	324	121	64	0

μ_1, μ_2, μ_3 et μ_4 ne changent pas ==> convergence

   (b) On aimerait maintenant comparer la qualité de ces regroupements. Pour cela, on recommence par regarder l'inertie intra-cluster.
      i. Calculer cette valeur pour les 3 regroupements précédents.

$J_w(k = 2) = 9 + 4 + 25 + 16 + 16 = 70$
$J_w(k = 3) = 0.25 + 0.25 + 2.25 + 2.25 + 0 = 5$
$J_w(k = 4) = 0.25 + 0.25 + 0 + 0 + 0 = 0.5$

      ii. En utilisant ce critère, quel serait le meilleur regroupement possible ? est-ce que cela vous parait réaliste ?

En poursuivant le raisonnement, avec un regroupement en 5 clusters (1 par point), on obtient $J_w(k = 5) = 0$. Ce critère ne nous permet donc pas de trouver le meilleur regroupement possible.

(c) S'inspirant du critère BIC, quelqu'un propose de rajouter le terme suivant au critère précédent : $+2kN \log N$ (où N est le nombre de données).

i. Expliquer l'utilité de ce terme

ce terme permet de jouer sur la complexité du modèle (principe du rasoir d'Occam), i.e. trouver le modèle qui regroupe le mieux les points, mais en evitant de "sur-apprendre", i.e. arriver à un modèle pour lequel 1 cluster = 1 point.

ii. Calculer la valeur du nouveau critère pour vos 3 regroupements. Qu'en concluez-vous

k	2	3	4	5
$J_w(k)$	70	5	0.5	0
$2kN \log N$	32.1888	48.2831	64.3775	80.4719
J+pen.	102.1888	53.2831	64.8775	80.4719

En cherchant à minimiser l'inertie intra-cluster pénalisée par la complexité du modèle, on trouve que le meilleur regroupement possible est celui correspondant à $k = 2$.

2. (6 points) Classification Hiérarchique

(a) Classification Hiérarchique Ascendante

i. Appliquer l'algorithme de classification hiérarchique ascendante en utilisant le saut minimal et tracer le dendrogramme correspondant.

	1	2	9	12	20
1		1	8	11	19
2			7	10	18
9				3	11
12					8

Regroupement des clusters {1} et {2} en {1+2}

	1+2	9	12	20
1+2		7	10	18
9			3	11
12				8

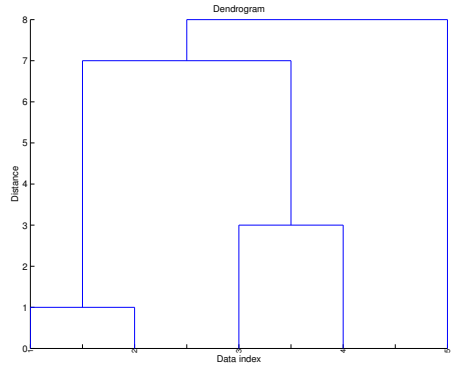
Regroupement des clusters {9} et {12} en {9+12}

	1+2	9+12	20
1+2		7	18
9+12			8

Regroupement des clusters {1+2} et {9+12} en {1+2+9+12}

	1+2+9+12	20
1+2+9+12		8

Regroupement des clusters {1+2+9+12} et {20} en {1+2+9+12+20}



ii. Idem avec le saut maximal.

	1	2	9	12	20
1		1	8	11	19
2			7	10	18
9				3	11
12					8

Regroupement des clusters {1} et {2} en {1+2}

	1+2	9	12	20
1+2		8	11	19
9			3	11
12				8

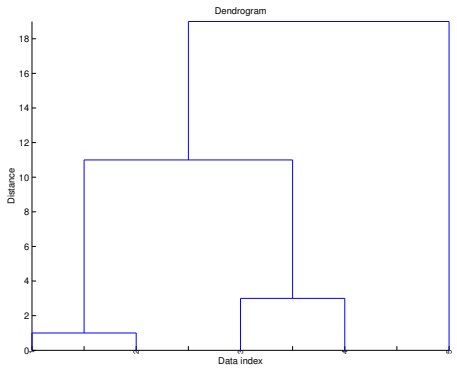
Regroupement des clusters {9} et {12} en {9+12}

	1+2	9+12	20
1+2		11	19
9+12			11

Regroupement des clusters {1+2} et {9+12} en {1+2+9+12}

	1+2+9+12	20
1+2+9+12		19

Regroupement des clusters {1+2+9+12} et {20} en {1+2+9+12+20}



(b) Classification Hiérarchique Descendante

Soit un algorithme de classification hiérarchique descendante qui recherche à chaque itération la meilleure façon de couper un ensemble de points en deux parties

i. Détailler la première itération de cet algorithme (en utilisant un saut minimal)

Il faut chercher comment couper le mieux possible l'ensemble des 5 points (i.e. avec la plus grande distance entre les 2 clusters formés)

C1	C2	dist(C1,C2)
1	2+9+12+20	1
2	1+9+12+20	1
9	1+2+12+20	7
12	1+2+9+20	3
20	1+2+9+12	8
1+2	9+12+20	7
1+9	2+12+20	1
1+12	2+9+20	1
1+20	2+9+12	1
2+9	1+12+20	1
2+12	1+9+20	1
2+20	1+9+12	1
9+12	1+2+20	7
9+20	1+2+12	3
12+20	1+2+9	3

Maintenant il faut faire pareil avec {20} (c'est fini) et avec {1+2+9+12}...

ii. Expliquer l'utilité de cet algorithme.

A chaque itération, la recherche de la meilleure séparation en 2 clusters est exponentielle, il faut parcourir tous les sous-ensembles possibles ... complexité algorithmique trop lourde en grande dimension