Invatare Automata

Laborator 2

Clasificare ierarhica

In cazul gruparii ierarhice nu este necesar sa pornim cu numarul de clustere fixat.

Clasificarea ierarhica are doua variante [1]:

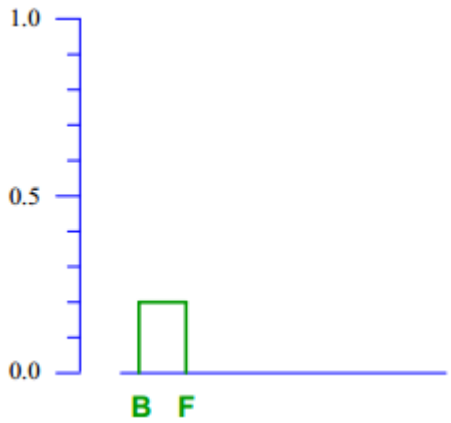
- Clasificare aglomerativa:
  - fiecare valoare de intrare apartine unui cluster.
  - Repeta
    - Unifica cele mai apropiate doua clustere
  - Pana cand exista un singur cluster

- Clasificare prin divizare:
  - Toate valorile de intrare fac parte din acelasi cluster
  - La fiecare pas se alege un cluster pentru a fi divizat

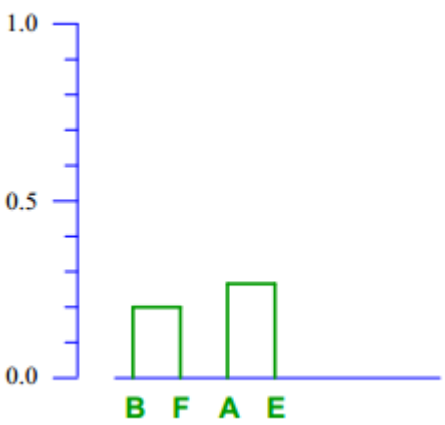
Rezultatul produs de clasificarea ierarhica este un arbore. Arborele pentru reprezentarea ierahica poarta numele de dendrograma (dimensiunile de-a lungul axei oy sunt reprezentate pe baza distantelor inter-clustere).

Exemplu: [2]

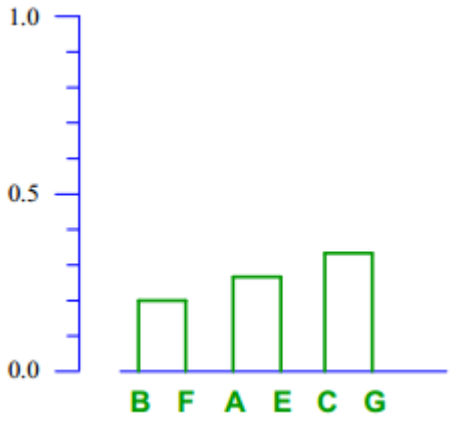
samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0



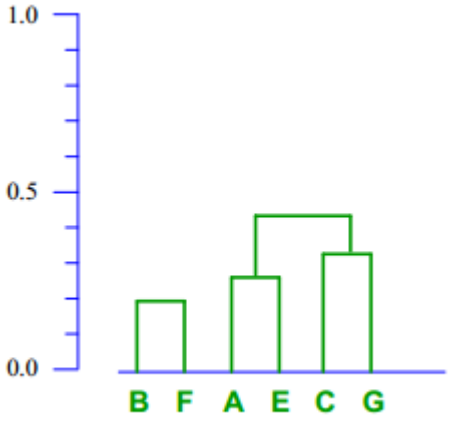
samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0



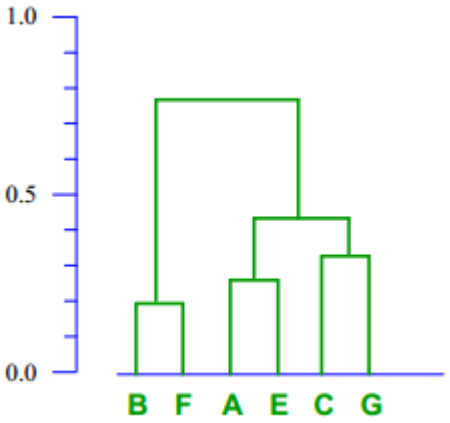
samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0



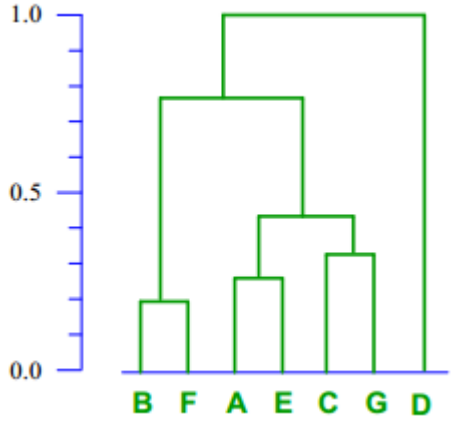
samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0



samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

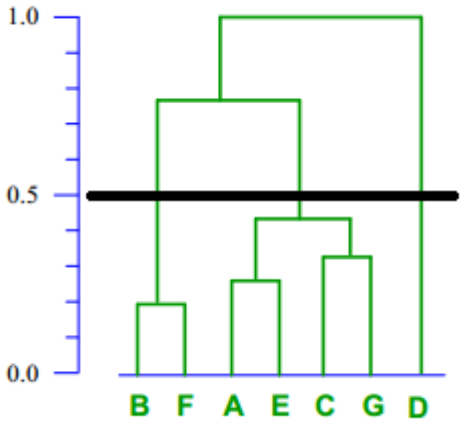


samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0



Alegerea numarului de clustere: se taie dendrograma cu o dreapta D, paralela cu axa ox

Exemplu: D = 0.5



clusterele formate vor fi:

C1: (B, F)
C2: (A, E, C, G)
C3: (D)

Similaritatea intre clustere (distanta inter-clustere):

- Single-linkage clustering: $d(A,B) = \min_{i \in A,\, j \in B} d_{ij}$
- Complete-linkage clustering: $d(A,B) = \max_{i \in A,\, j \in B} d_{ij}$
- Average linkage clustering: $d(A,B) = \frac{1}{|A| * |B|} \sum_{i \in A} \sum_{j \in B} d_{ij}$

Resurse

[1] http://en.wikipedia.org/wiki/Hierarchical_clustering
[2] Hierarchical cluster analysis : http://www.econ.upf.edu/~michael/stanford/maeb7.pdf