

# Învățare Automată - Laboratorul 2

## Grupare Ierarhică

Tudor Berariu

*tudor.berariu@gmail.com*

Laboratorul AIMAS

Facultatea de Automatică și Calculatoare

Universitatea Politehnica București

28 februarie 2016

## 1 Scopul laboratorului

Scopul laboratorului îl reprezintă înțelegerea grupării ierarhice și a diferențelor dintre aceasta și algoritmul **K-Means** prezentat în primul laborator.

## 2 Introducere

În primul laborator a fost prezentat un algoritm pentru grupare, **K-Means**, care suferea de câteva limitări importante (soluția depindea de alegerea centroizilor inițiali; numărul de grupuri trebuie cunoscut sau intuit). De asemenea, s-a observat faptul că anumite seturi de date nu pot fi grupate folosind algoritmul **K-Means** din cauza *forme* grupurilor. Pentru a rezolva acele seturi de date este nevoie de o abordare diferită de cea a reprezentării grupurilor prin centroizi. O altă abordare a problemei grupării unei mulțimi de obiecte în funcție de similaritatea dintre acestea (engl. *cluster analysis*) o reprezintă gruparea ierarhică (engl. *hierarchical clustering*).

Spre deosebire de algoritmul **K-Means**, în cazul grupării ierarhice nu este necesară stabilirea a priori a numărului de grupuri și a unei partiționări inițiale a obiectelor.

Gruparea ierarhică are două variante:

**grupare aglomerativă** în care se pornește de la situația în care fiecare obiect formează singur un grup. Apoi se reunesc succesiv cele mai apropiate două grupuri până când rămâne unul singur.

**grupare prin divizare** în care se pornește de la un singur grup ce cuprinde toate obiectele, iar la fiecare pas se alege un grup (cel mai eterogen) pentru a fi segmentat.

Rezultatul produs de gruparea ierarhică este un arbore de grupări / divizări succesive. De obicei, acest arbore este reprezentat grafic, pentru proporții ținându-se cont de similaritatea

### 3 Măsurarea apropierii dintre două grupuri

Gruparea ierarhică reprezintă, de fapt, o familie de algoritmi ce folosesc diferite definiții ale distanței (similarității) dintre obiecte pentru construirea clusterelor.

**single-linkage** - distanța (similaritatea) dintre cele mai apropiate două puncte

**complete-linkage** - distanța (similaritatea) dintre cele mai depărtate două puncte

**group-average** - distanța (similaritatea) medie a celor două grupuri

2

## 4 Alte metode de grupare

În afara metodelor bazate pe centroid (**K-Means**) și pe conectivitate (gruparea ierarhică), alte modele de grupare sunt: metode bazate pe distribuții statistice, metode bazate pe densitate (DBSCAN, OPTICS), biclustering.

## 5 Cerințe

În cadrul acestui laborator trebuie rezolvate cerințele de mai jos. Scheletul de cod conține funcții pentru citirea datelor, afișarea dendrogramei și a setului de date colorat conform grupării create.

1. [6 puncte] Implementați într-un limbaj de programare la alegere algoritmul de grupare ierarhică aglomerativă folosind distanța *single-linkage*. Implementați funcția **singleLinkage** ce trebuie să întoarcă o matrice cu  $(N - 1) \times 4$  valori. Fiecare linie corespunde unei alipiri a două grupuri (plecând de la  $N$  grupuri se ajunge la unul singur în  $N - 1$  pași). Primele două valori corespund id-urilor grupurilor ce trebuie unite, a treia valoare conține distanța dintre cele două grupuri, iar cea de-a patra numărului de puncte pe care le conține noul grup. Pentru id-uri, valorile de la 0 la  $N - 1$  se referă la exemplele din setul de date, iar cele de la  $N$  la  $2N - 2$  corespund grupurilor construite pe parcurs (la pasul  $0 \leq i < N - 1$  se construiește clusterul  $N + i$ ).
2. [2 puncte] Implementați și celelalte două variante de algoritmi: utilizând distanța *complete-linkage* și *group-average*. Cele două funcții întorc o matrice cu aceeași semantică precum în cazul *single-linkage*.
3. [2 puncte] Implementați funcția **extractClusters** care pe baza unei aglomerări ierarhice construite anterior, stabilește numărul optim de clustere ca fiind cel dinaintea alipirii făcute la cea mai mare distanță.
4. [2 puncte] Testați algoritmul implementat și eficiența acestuia pe seturile de date din arhivă. O descriere a acestora se găsește în Anexa A. Comparați pentru câteva seturi de date acuratețea celor trei metode și comparați-le între ele, dar și cu algoritmul K-Means din laboratorul 1.

## A Seturi de date

În cadrul acestui laborator veți folosi seturile de date FCPS<sup>1</sup> (Fundamental Clustering Problem Suite) ale Philipps Universität Marburg. Acestea se găsesc în arhiva **FCPS.zip**.

Pentru fiecare set de date veți găsi următoarele fișiere în subdirectorul **01FCPSdata**:

- **<nume>.lrn** - setul de date cu un id pentru fiecare obiect,
- **<nume>.cls** - clasele *reale* ale obiectelor.

---

<sup>1</sup><http://www.uni-marburg.de/fb12/datenbionik/downloads/FCPS>

Coloanele sunt separate prin TAB.

De asemenea în directorul `02Documentation` se găsesc reprezentări grafice ale seturilor de date.