Învățare Automată - Laboratorul 1
Algoritmul K-Means

Tudor Berariu
Laboratorul AIMAS
Facultatea de Automatică și Calculatoare

17 februarie 2014

1 Scopul laboratorului

Scopul laboratorului îl reprezintă înțelegerea și implementarea unui algoritm de învățare nesupervizată pentru grupare: **K-Means**.

2 Introducere

Una dintre problemele fundamentale ale învățării automate o reprezintă alocarea unui set de obiecte unor grupuri (eng. *clusters*) astfel încât obiectele din același grup să prezinte un grad mare de similaritate. Această problemă de învățare nesupervizată se numește *cluster analysis*

3 Algoritmul K-Means

Algoritmul **K-Means** [M$^+$67] pornește de la un set de k centroizi aleși la întâmplare din setul de obiecte. Se repetă alternativ următorii doi pași până când algoritmul *converge*:

1. Se parcurg toate obiectele din setul de date și fiecare dintre acestea este alocat grupului corespunzător celui mai apropiat centroid.
2. Se recalculează centroidul fiecărui grup.

Algorithm 1 K-Means

1: se aleg la întâmplare k centroizi: $c_1, \ldots, c_k$
2: **repeat**
3:
4: **for** $i = 1$ **to** k **do**
5: $C_i \leftarrow \{x \in \mathbf{X}, d(x, c_i) < d(x, c_j)\ \forall j \neq i\}$
6: **end for**
7: **for** $i = 1$ **to** k **do**
8: $c_i \leftarrow \frac{1}{|C_i|} \sum_{x \in C_i} x$
9: **end for**
10: **until** algoritmul converge

Algoritmul se oprește atunci când în urma unei iterații nu s-a modificat componența grupurilor.

4 Limitări ale algoritmului K-Means

Algoritmul K-Means prezintă trei probleme importante:

1. numărul de grupuri k trebuie cunoscut a priori;
2. algoritmul converge către un minim local;
3. rezultatul algoritmuui depinde de alegerea centroizilor inițiali.

5 Alegerea centroizilor inițiali

În algoritmul clasic K-Means cei k centroizi inițiali se aleg aleator din mulțimea obiectelor din setul de date. În continuare sunt descrise două rețete mai bune pentru acest pas.

5.1 Algoritmul K-Means++

Algoritmul **K-Means++** [AV07] reprezintă o variantă îmbunătățită a algoritmului K-Means în care centroizii inițiali sunt aleși după cum urmează. Primul centroid se alege aleator din setul de date. Următorii $k - 1$ se aleg

succesiv dintre obiectele din setul de date cu o probabilitate $\frac{D(x_i)^2}{\sum_{x \in \mathbf{X}} D(x)^2}$ pentru fiecare obiect $x_i \in \mathbf{X}$. În formula precedentă $D(x)$ este distanța cea mai mică dintre obiectul x și un centroid deja ales.

5.2 Metoda Kaufman

In [PLL99] s-au testat pe diferite seturi de date mai multe metode de inițalizare a centroizilor pentru algoritmul K-Means. Rezultatele au arătat că una dintre cele mai bune metode este cea propusă de Kaufman. Se alege întâi cel mai central obiect din setul de date, iar apoi se adaugă succesiv acele obiecte care strâng în jurul lor cel mai mare număr de elemente (vezi Algoritmul 2).

Algorithm 2 Algoritmul Kaufman pentru alegerea centroizilor inițiali

$s_1 \leftarrow$ cel mai central obiect din X
for a=2 **to** k **do**

  for x_i obiect neselectat **do**

    for x_j obiect neselectat, $x_i \neq x_j$ **do**
      $D_j \leftarrow \min_l d(s_l, x_j)$
      $C_{ij} \leftarrow \max(D_j - d(x_i, x_j)), 0)$
    end for
    $G_i \leftarrow \sum_j C_{ij}$
  end for
  $s_a \leftarrow x_i$ pentru care G_i exte maxim
end for

6 Cerințe

În cadrul acestui laborator trebuie rezolvate următoarele cerințe:

1. [8 puncte] Implementați într-un limbaj de programare la alegere algoritmul K-Means descris în Secțiunea 3.

2. [2 puncte] Testați algoritmul implementat și eficiența acestuia pe seturile de date din arhivă. O descriere a acestora se găsește în Anexa A.

3. [2 puncte] Implementați unul dintre cei doi algoritmi prezentați în Secțiunea 5:

   - metoda Kaufman pentru alegerea centroizilor inițiali sau
   - algoritmul K-Means++.

A Seturi de date

În cadrul acestui laborator veți folosi seturile de date FCPS [1] (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.

Coloanele sunt separate prin TAB.

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

Bibliografie

[AV07] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[M+67] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

[1] `http://www.uni-marburg.de/fb12/datenbionik/downloads/FCPS`

[PLL99] José Manuel Pena, Jose Antonio Lozano, and Pedro Larranaga. An empirical comparison of four initialization methods for the¡ i¿ k¡/i¿-means algorithm. *Pattern recognition letters*, 20(10):1027–1040, 1999.