Învățare Automată - Laboratorul 1
Grupare. Algoritmul K-Means

Tudor Berariu
tudor.berariu@gmail.com
Laboratorul AIMAS
Facultatea de Automatică și Calculatoare
Universitatea Politehnica București

23 februarie 2016

1 Scopul laboratorului

Scopul laboratorului îl reprezintă înțelegerea și implementarea unei metode de învățare nesupervizată pentru grupare (engl. *clustering*): algoritmul **K-Means**.

Structura documentului este următoarea. Secțiunea 2 prezintă contextul teoretic și formalizează problema ce se dorește rezolvată. Secțiunea 3 descrie în detaliu algoritmul **K-Means**. Secțiunea 4 enumeră câteva dintre limitările algoritmului **K-Means** și oferă câteva soluții simple pentru depășirea acestora. Una dintre aceste probleme se referă la importanța centroizilor inițiali, iar Secțiunea 5 descrie câteva metode avansate pentru alegerea acestora.

Secțiunea 7 conține cerințele ce trebuie rezolvate în cadrul laboratorului, datele folosite pentru testare fiind descrise în Anexa B.

2 Problema

Una dintre problemele fundamentale ale învățării automate o reprezintă identificarea grupurilor (engl. *clusters*) într-un set de obiecte astfel încât obiectele din același grup să prezinte un grad mare de similaritate. Această problemă de învățare nesupervizată se numește *cluster analysis*.

Problema grupării se poate formaliza în diferite feluri, existând mai multe abordări. În acest laborator vom rezolva problema grupării bazate pe centroizi (engl. *centroid-based clustering*).

Se consideră un set de date $\mathbf{X} = \{\mathbf{x}_1, \dots \mathbf{x}_N\}$ ce conține N exemple într-un spațiu D-dimensional. Scopul este obținerea unei partiții a setului de date în K grupuri reprezentate prin K vectori prototip $\mathbf{c}_k, 1 \leq k \leq K$, numiți *centre* sau *centroizi*, astfel încât distanța

totală de la fiecare exemplu la cel mai apropiat centroid (Formula 1) să fie minimă.

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} in_{i,k} \cdot ||\mathbf{x}_i - \mathbf{c}_k||^2 \tag{1}$$

$$in_{i,k} = \begin{cases} 1 & \text{daca } k = \underset{l}{\text{argmin}} ||\mathbf{x}_i - \mathbf{c}_l||^2 \\ 0 & \text{altfel} \end{cases} \tag{2}$$

3 Algoritmul K-Means

Algoritmul **K-Means** [M+67] pornește de la un set de K centroizi aleși aleator 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.

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

Algoritmul 1 Algoritmul **K-Means**

funcția K-Means$(\mathbf{X}, \mathbf{K})$
  pentru $k \leftarrow 1 \dots K$ **execută**
    $\mathbf{c}_k \leftarrow rand(\mathbf{X})$ ▷ centroizii inițiali se aleg aleator din setul de date
  repetă
    pentru $i \leftarrow 1 \dots N$ **execută**
      $in_{i,k} \leftarrow \begin{cases} 1 & \text{daca } k = \underset{l}{\text{argmin}} ||\mathbf{x}_i - \mathbf{c}_l||^2 \\ 0 & \text{altfel} \end{cases}$ ▷ exemplul este asociat celui mai apropiat grup
    pentru $k \leftarrow 1 \dots K$ **execută**
      $\mathbf{c}_k \leftarrow \frac{\sum_{i=1}^{N} in_{i,k} \mathbf{x}_i}{\sum_{i=1}^{N} in_{i,k}}$ ▷ $\mathbf{c}_k$ devine media tuturor exemplelor din grupul k
  până când algoritmul converge

Pentru a înțelege de ce algoritmul **K-Means** găsește un minim al expresiei J (Formula 1), trebuie observat că cei doi pași optimizează succesiv parametrii $in_{i,k}$ și $\mathbf{c}_k$.

1. Fixând $\mathbf{c}_k$, se recalculează $in_{i,k}$ conform ecuației 2.

2. Fixând $in_{i,k}$, un minim al expresiei J se găsește în punctul în care derivata este zero:

$$\sum_{i=1}^{N} in_{i,k}(\mathbf{x}_i - \mathbf{c}_k) = 0 \tag{3}$$

4 Limitări ale algoritmului K-Means

Algoritmul **K-Means** prezintă câteva limitări importante.

1. Numărul de grupuri K trebuie cunoscut a priori.
   - Dacă acest număr nu este cunoscut, se poate rula algoritmul pentru diferite valori ale lui K și se poate alege o partiție convenabilă. Altfel, se poate alege o altă metodă de grupare, de exemplu o strategie de grupare ierarhică.
2. Algoritmul converge către un minim local.
   - Nu există o metodă tractabilă care să garanteze un minim global. În practică se obișnuiește rularea algoritmului de mai multe ori și păstrarea celui mai bun rezultat.
3. Rezultatul algoritmului depinde de alegerea centroizilor inițiali.
   - Există mai multe strategii pentru alegerea centroizilor inițiali ($\mathbf{c}_k, 1 \leq k \leq K$), două dintre acestea fiind descrise în Secțiunea 5.

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 $\mathbf{c}_1$ se alege aleator din setul de date. Următorii $K-1$ se aleg succesiv dintre obiectele din setul de date cu o probabilitate

$$p_i = \frac{D(\mathbf{x}_i)^2}{\sum_{\mathbf{x} \in \mathbf{X}} D(\mathbf{x})^2}$$

pentru fiecare obiect $\mathbf{x}_i \in \mathbf{X}$, unde $D(\mathbf{x})$ este distanța cea mai mică dintre obiectul $\mathbf{x}$ 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.

Algoritmul 2 Algoritmul Kaufman pentru alegerea centroizilor inițiali

funcția KAUFMAN($\mathbf{X}, \mathbf{K}$)
  $\mathbf{c}_1 \leftarrow$ cel mai central exemplu din $\mathbf{X}$
  pentru $k \leftarrow 2 \dots K$ **execută**
    pentru $\mathbf{x}_i \in \mathbf{X}$ **execută**
      pentru $\mathbf{x}_j \in \mathbf{X} \setminus \{\mathbf{x}\}$ **execută**
        $d_j \leftarrow \min_{l \in 1 \dots k-1} ||\mathbf{c}_l - \mathbf{x}_j||$
        $c_{ij} \leftarrow \max(d_j - ||\mathbf{x}_i - \mathbf{x}_j||, 0)$
      $g_i \leftarrow \sum_j c_{ij}$
    $idx \leftarrow \operatorname{argmax}_i g_i$
    $\mathbf{c}_k \leftarrow \mathbf{x}_{idx}$ ▷ se alege $\mathbf{x}$ pentru care G este maxim

6 Evaluarea unei grupări

Nu există o rețetă unică pentru evaluarea unei grupări realizate pentru un set de date. În general, metricile țin cont de faptul că exemplele dintr-un grup trebuie să fie cât mai apropiate / similare, iar cele din grupuri diferite trebuie să fie cât mai diferite.

Dacă sunt cunoscute clasele *reale* (precum într-o problemă de învățare supervizată), atunci evaluarea se poate face mai ușor. Dintre metodele existente, este descrisă în continuare *Rand Index*.

6.1 Rand Index

Fiind date o grupare C și valorile reale T ale claselor din care fac parte obiectele dintr-un set de date, definim:

TP numărul de perechi i, j care sunt în același grup în C și au aceeași clasă în T;

FP numărul de perechi i, j care sunt în același grup în C, dar sunt în clase diferite în T;

FN numărul de perechi i, j care sunt în grupuri diferite în C, dar au aceeași clasă în T;

TN numărul de perechi i, j care sunt în grupuri diferite în C și au clase diferite în T.

Metrica *Rand Index* este definită în Formula 4:

$$R \triangleq \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

Evident, $0 \leq R \leq 1$.

7 Cerințe

În cadrul acestui laborator trebuie rezolvate următoarele cerințe:

1. [6 puncte] Implementați într-un limbaj de programare la alegere algoritmul **K-Means** descris în Secțiunea 3.
2. [2 puncte] Implementați metrica *Rand Index* descrisă în Secțiunea 6.1.
3. [2 puncte] Testați algoritmul implementat și eficiența acestuia pe seturile de date din arhivă. O descriere a acestora se găsește în Anexa B. Explicați de ce pe unele seturi de date rezultatele sunt *nesatisfăcătoare*.
4. [2 puncte] Implementați unul dintre cei doi algoritmi prezentați în Secțiunea 5:
   - metoda Kaufman pentru alegerea centroizilor inițiali sau
   - algoritmul K-Means++.

   Comparați grupările obținute astfel cu cele obținute cu algoritmul **K-Means**. Este utilă inițializarea atentă a centroizilor?

Puteți folosi funcțiile pentru citirea datelor și pentru afișarea grafică a unei grupări din scheletul de cod (descris în Anexa **??**).

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.

B Scheletul de cod

Funcția `getArchive` descarcă arhiva `FCPS.zip` dacă aceasta nu este în directorul curent și întoarce un obiect de tipul `ZipFile`.

Funcția `getDataSet` primește obiectul de tip `ZipFile` și numele setului de date dorit (e.g. *Atom*) și întoarce o matrice `X` cu datele și un vector `T` cu clasele adevărate.

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

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.

[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.