Telecom Bretagne – 2016-2017

Département Logique des Usages, Sciences Sociales et de l'Information
Patrick Meyer, Philippe Lenca, Sorin Moga
sorin.moga@telecom-bretagne.eu

INF 413 – **Algorithmique avancée**

TP : Etude et implémentation de l'algorithme des `k-means`.
Illustration d'algorithmes dont le résultat et la complexité dépendent de choix réalisés par l'utilisateur.
ENONCÉ

Rapprocher ce qui se ressemble et éloigner ce qui diffère

Une tâche fréquente en analyse de données consiste, à partir d'un ensemble d'observations à créer des groupes d'individus de telle sorte que les individus d'un groupe donné aient tendance à être similaires, et en même temps aient tendance à être différents des individus des autres groupes. Les algorithmes de classification répondent à cette tâche.

L'objet de ces travaux pratiques consiste à étudier, implémenter, tester et discuter l'algorithme `k-means` l'un des algorithmes de classification à la fois parmi les plus populaires et les plus simples.

Une brève description de l'algorithme ainsi qu'un exemple jouet sont donnés ci-après. On vous demande d'étudier l'algorithme de façon la plus autonome possible (en recherchant des présentations et des discussions de l'algorithme sur internet par exemple ou lire l'article de Tapas Kanungo et al. distribué avec ce sujet de travaux pratiques (Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu: An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24(7): 881-892 (2002)).

Vous rendrez un compte rendu qui servira d'évaluation de type *contrôle continu* (veillez à citer correctement vos sources!).

Principales étapes de l'algorithme

- initialisation de k groupes $\mathtt{G_i}$: choisir k individus $\mathtt{o_i}$ formant ainsi k centres $\mathtt{c_i}$ (par exemple avec une fonction aléatoire)
- affecter les individus au groupe dont ils sont le plus proche : (ré)affecter chaque individu $\mathtt{o}$ au groupe $\mathtt{G_i}$ de centre $\mathtt{c_i}$ tel que $\mathtt{dist}(\mathtt{o}, \mathtt{c_i})$ est minimal (par exemple avec la distance Euclidienne)
- mettre à jour le représentant des groupes : recalculer $\mathtt{c_i}$ de chaque groupe (par exemple en calculant le barycentre)
- aller à l'étape 2 selon le critère d'arrêt (par exemple un nombre maximal d'itérations, ou si aucun individu à changé de groupe)

Exemple

Déroulons l'algorithme `k-means` sur les données de la table 1 avec $k = 3$ (les centres initiaux étant d, k et m) et la distance Euclidienne.

o_i	v_1	v_2	v_3	v_4	v_5
a	7	8	4	5	2
b	6	8	5	4	2
c	8	9	7	8	9
d	6	7	7	7	8
e	1	2	5	3	4
f	3	4	5	3	5
g	7	8	8	6	6
h	8	9	6	5	5
i	2	3	5	6	5
j	1	2	4	4	2
k	3	2	6	5	7
l	2	5	6	8	9
m	3	5	4	6	3
n	3	5	5	6	3

Table 1: Ensemble d'individus

Pour chaque individu de la table 1 on calcule ses distances aux trois centres et on l'affecte au groupe du centre dont il est le plus proche. On obtient la table 2.

o_i	$\texttt{dist}(\texttt{o}_\texttt{i}, \texttt{c}_1 = \texttt{d})$	$\texttt{dist}(\texttt{o}_\texttt{i}, \texttt{c}_2 = \texttt{k})$	$\texttt{dist}(\texttt{o}_\texttt{i}, \texttt{c}_3 = \texttt{m})$	G_i
a	7,1	9,0	5,2	3
b	7,1	8,5	4,9	3
c	3,2	9,4	9,5	1
e	9,3	4,2	4,9	2
f	6,9	3,6	3,9	2
g	2,8	7,6	7,1	1
h	4,7	8,8	7,1	1
i	6,8	2,8	3,2	2
j	10,2	5,8	4,2	3
l	4,8	4,8	6,7	1
n	6,6	5,2	1,0	3

Table 2: Distance de chaque individu aux centres et affectation aux groupes

Ensuite, le centre de chaque groupe doit être mis à jour : tables 3, 4 et 5.

o_i	v_1	v_2	v_3	v_4	v_5
c	8	9	7	8	9
d	6	7	7	7	8
g	7	8	8	6	6
h	8	9	6	5	5
l	2	5	6	8	9
c_1	6,2	7,6	6,8	6,8	7,4

Table 3: Mise à jour de c_1

On continue le même processus jusqu'au critère d'arrêt. Chaque individu est comparé aux nouveaux centres et affecté au groupe dont il est le plus proche, etc., si un individu est déplacé d'un groupe à un autre alors les deux centres seront modifiés.

o_i	v_1	v_2	v_3	v_4	v_5
e	1	2	5	3	4
f	3	4	5	3	5
i	2	3	5	6	5
k	3	2	6	5	7
c_2	2,25	2,75	5,25	4,25	5,25

Table 4: Mise à jour de c_2

o_i	v_1	v_2	v_3	v_4	v_5
a	7	8	4	5	2
b	6	8	5	4	2
j	1	2	4	4	2
m	3	5	4	6	3
n	3	5	5	6	3
c_3	4,0	5,6	4,4	5,0	2,4

Table 5: Mise à jour de c_3

Directives pour le TP

Votre travail se divise en 3 parties:

- Etude des k-means et des différents points clés de l'algorithme;
- Implémentation de l'algorithme et des points étudiés en Python;
- Rédaction d'un rapport sur l'étude et présentation de l'algorithme sous la forme de pseudo-code et étude de la complexité de l'algorithme.

L'étude consistera à comprendre l'algorithme et à détecter les différents endroits de l'algorithme pour lesquels plusieurs options se présentent.

Pour l'implémentation, vous veillerez à respecter les quelques indications ci-dessous:

- Décomposez votre code en plusieurs fonctions:
  - Génération aléatoire de données en fonction du nombre d'attributs et du nombre d'observations;
  - Génération aléatoire de centres;

- Chargement de données à partir de fichiers et sauvegarde des données et des résultats dans des fichiers;
- Calcul de la distance;
- Affectation des observations aux classes;
- Condition d'arrêt;
- ...

- Documentez bien chacune des fonctions, rajoutez des commentaires dans le code et documentez l'utilisation de votre programme;
- En sortie, fournissez un fichier contenant les données avec leur classe correspondante, ainsi qu'un fichier contenant les coordonnées des centres.

Les formats suivants de fichiers sont à respecter:

Le fichier des centres

```
# no_centre,attribut_1,attribut_2,...,attribut_p
1,0.2,0.3,0.4,0.5
2,0.3,0.2,0.6,0.4
3,0.1,0.9,0.5,0.3
```

Le fichier des données d'entrée

```
# no_observation,attribut_1,attribut_2,...,attribut_p
1,0.1,0.8,0.4,0.3
2,0.3,0.7,0.5,0.2
3,0.9,0.6,0.7,0.1
4,0.4,0.2,0.8,0.3
```

Le fichier des données affectées

```
# no_observation,attribut_1,attribut_2,...,attribut_p,no_classe
1,0.1,0.8,0.4,0.3,3
2,0.3,0.7,0.5,0.2,2
3,0.9,0.6,0.7,0.1,1
4,0.4,0.2,0.8,0.3,3
```

Evaluation

A la fin de chaque séance (le 13 mars 2017 avant 14h et le 24 mars 2017 avant 14h), veuillez à déposer sur moodle une archive contenant l'état d'avancement de vos travaux. Le nom de fichier doit être *Nom_Nom_INF413_2016_2017_P.zip*

Le rapport final, ainsi que les sources Python et le fichier README (qui décrit l'utilisation de votre programme), sont à déposer dans une archive sur moodle pour le 15 avril 2017 à minuit. Le nom de fichier doit être *Nom_Nom_INF413_2016_2017_P.zip*.

Bonus

Ceux qui veulent aller plus loin, peuvent étudier et implémenter l'algorithme G-means, une version de K-means pour laquelle on apprends aussi le K. Une brève description de l'algorithme est donné dans l'article *HAMERLY, Greg et ELKAN, Charles. Learning the k in k-means. Advances in neural information processing systems, 2004, vol. 16, p. 281* disponible sous moodle.