Université USTHB – Bab-Ezzouar Faculté d'Informatique, Département d'IASD, Module : Data Mining Bab-Ezzouar, 2022 / 2023 2^{ème} année Master Informatique Système Informatique Intelligent Semestre 1

Exercice 6 Algorithmes de Clustering

Distance

Distance euclidienne:	d(A,B) = $\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$
Distance de Manhattan :	$d(A, B) = \sum_{i=1}^{n} A_i - B_i $
Distance de Hamming :	$d(A, B) = \#\{i : A_i \neq B_i\}$

Algorithme

Algorithme: k-means

Entrée: D: Dataset; k: le nombre de cluster à former;

Sortie: D: Dataset étiqueté;

Début

Choisir aléatoirement k instances comme centroïdes.

Répéter

Calculer la distance entre chaque instances D[i] et les k centroïdes ;

Affecter chaque instances D[i] au groupe dont il est le plus proche de son centre;

Calculer le nouveau centre de chaque cluster et modifier le centroïde ;

Jusqu'à D[i] identique à D[i-1];

Retourner D;

Fin.

Questions:

- 1- Écrire une fonction java permettant de calculer la distance de Manhattan entre deux instances du dataset.
- 2- Écrire une fonction java permettant de calculer le centroïde d'un ensemble d'instances.
- 3- Écrire une fonction java permettant de trouver le cluster dont une instance donnée est la plus proche.
- 4- Implémenter en java l'algorithme k-means.
- 5- Déduire les clusters formés (sans considérer la classe comme un attribut).

HOUACINE Naila Aziza : n.a.houacine@gmail.com