M1 – Informatique

Université Paul Sabatier
31062 Toulouse Cedex 09

IAA – Introduction à l'Apprentissage Automatique

Travaux dirigés n°1 – Classification automatique

I Quantification vectorielle : une revisite de l'algorithme des « K means »

Objectif : réduire le nombre des observations en recherchant des prototypes. Un prototype pourra être assimilé à une classe. Le nombre de prototypes ou classes est fixé a priori, égal à K.

Soit l'ensemble de points suivants de R^2 :

Y={(1,1),(3,3),(1,5),(-3,5),(-5,3),(-3,1),(3,-1),(5,-3),(3,-5),(-1,-5),(-3,-3),(-1,-1)}

I.0 Positionner les points sur un graphe, l'ensemble de l'exercice peut être fait à partir de simples graphes, sans calcul important. A chaque fois, donner explicitement les dictionnaires optimaux.

I.1 Trouver les dictionnaires D de taille 2 à partir des dictionnaires initiaux suivants et en utilisant la méthode graphique :

1) D^0 = {(1,2), (-2,-1)}
2) D^0 = {(x,y), (-y,-x)} pour tout couple (x,y). Commenter ce résultat.
3) D^0 ={(3,3), (3,-2)}

Comparer les résultats et expliquer.

Pour le dictionnaire de 1), on appliquera également l'algorithme des k-means en utilisant le critère de recherche de $D=\{d_1, .., d_k\}$

$$Crit(D) = \sum_{n=1}^{N} d(y_n, d_{\hat{n}})^2$$

$$\hat{n} = \arg\min_{1 \le k \le K} d(y_n, d_k)$$

$$\hat{D} = \arg\min_{D} Crit(D)$$

I.2 Si l'on sait a priori que Y est composé de deux classes, comment choisir le dictionnaire initial pour s'assurer de la convergence de l'algorithme vers la partition de Y en ces deux classes ?

En l'occurrence Y est l'union de S1 et S2 avec :

S_1={(1,1),(3,3),(1,5),(-3,5),(-5,3),(-3,1)}

S_2={(3,-1),(5,-3),(3,-5),(-1,-5),(-3,-3),(-1,-1)}

I.3 Trouver le dictionnaire D de taille 3, construit à l'aide de l'algorithme des k-means, à partir du dictionnaire initial D^0 ={(3,1), (3,-2), (3,-6)}. Commenter.

II Classification hiérarchique non supervisée

On considère les 5 observations suivantes dans R :

$x_1 = 1, \quad x_2 = 7, \quad x_3 = 15, \quad x_4 = 5, \quad x_5 = 2$

Nous allons tester quatre méthodes pour trouver un ensemble de classes ; dans chaque cas donner la partition à 2 classes et celle à 3 classes. A la fin, comparer les résultats et les commenter.

II.1 Effectuer une classification hiérarchique ascendante (sous forme d'arbre indicé) par la méthode des distances. On utilisera la distance entre groupes suivante :

$d(i \cup j, k) = \min\{d(i,k), d(j,k)\}$ où i,j et k sont des groupes.

II.2 Que donne la méthode des K-means avec pour partition initiale : $\{x_1, x_2, x_3\} \quad et \quad \{x_4, x_5\}$

II.3 Effectuer une classification hiérarchique ascendante à l'aide de la méthode des moments d'ordre 2.

II.4 Effectuer une classification descendante avec l'algorithme de splitting (progressif, on ne perturbe que la classe de distorsion moyenne maximale, la perturbation ε est fixée à 1).