Algorithmes des k moyennes et EM

Maxime CHAMBREUIL
maxime.chambreuil@insa-rouen.fr

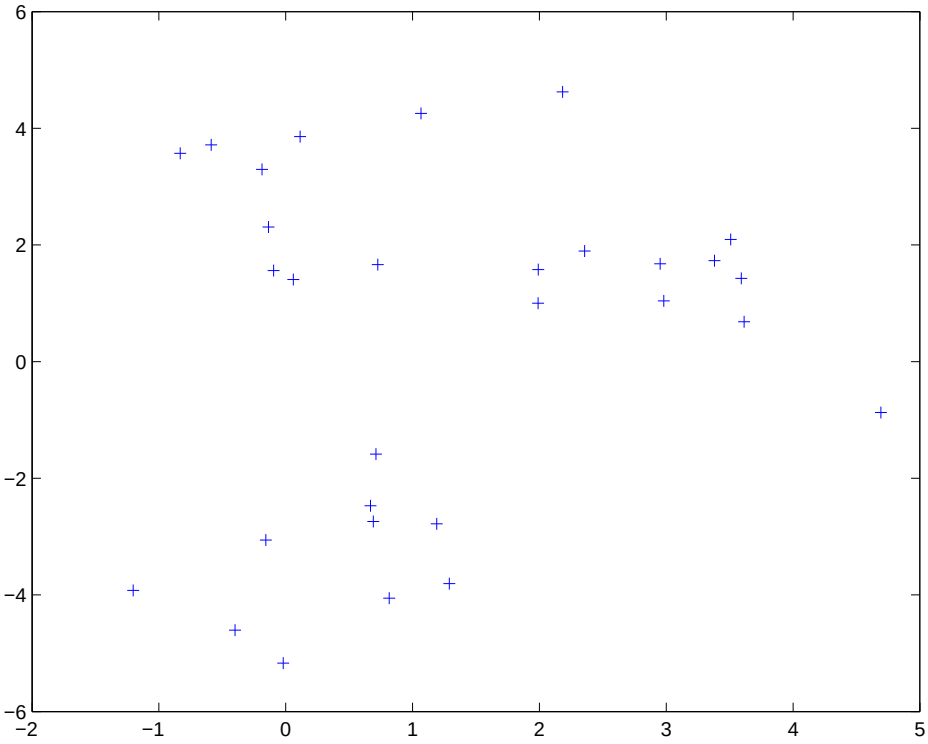
16 mars 2003

1 Explication

Nous avons généré un échantillon de 30 individus à partir de 3 lois normales, d'espérance différentes :

$$\mu_1 = (0, 3)$$
$$\mu_2 = (0, -3)$$
$$\mu_3 = (3, 1)$$

et de variance la matrice identité, pour obtenir :

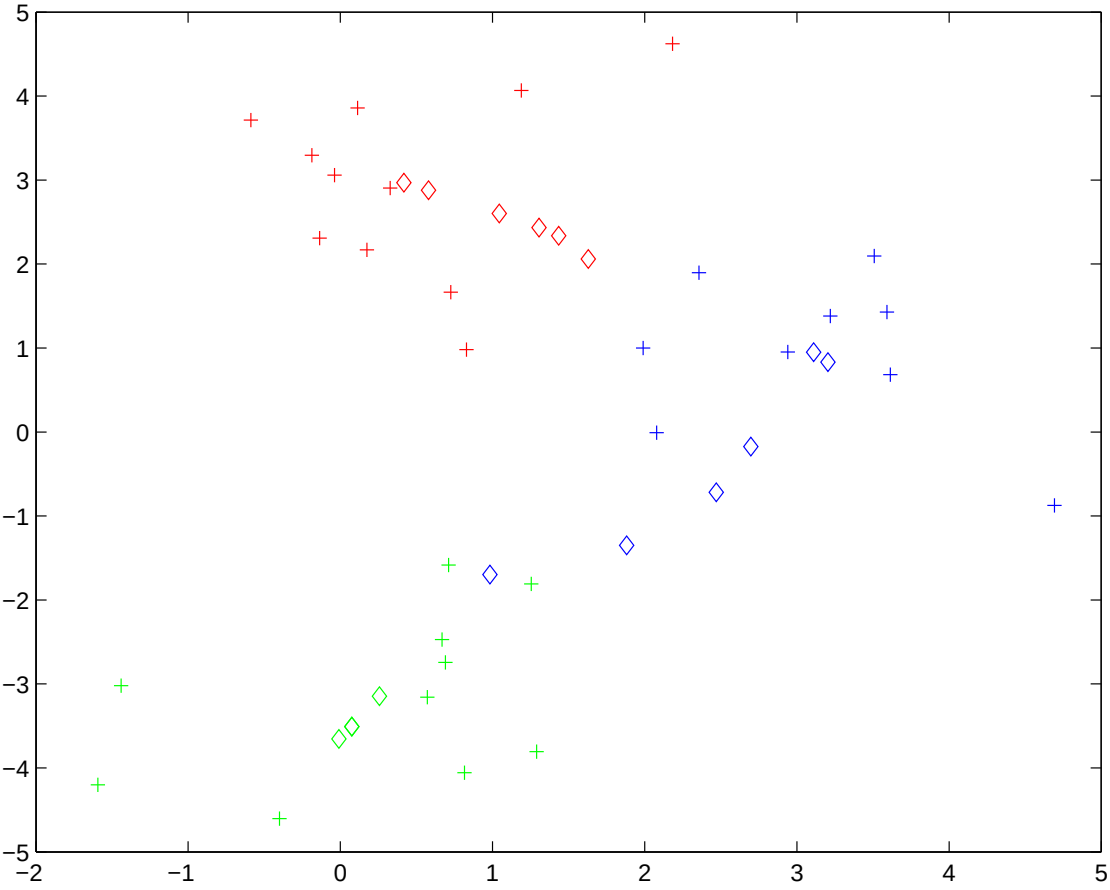


A côté de ça, nous avons généré 3 moyennes aléatoirement : m1, m2 et m3 ; correspondants aux classes C1, C2 et C3.

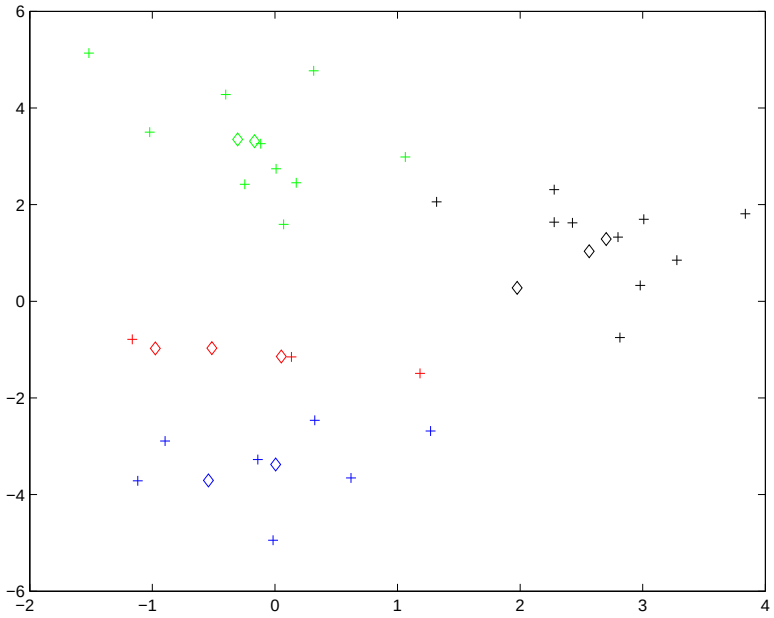
Ensuite, nous avons calculé la distance entre chaque individu et ses 3 moyennes. Nous avons classé chaque individu, en fonction de la distance minimale :

$$(x - m1) > (x - m2) > (x - m3) \Rightarrow x \in C3$$

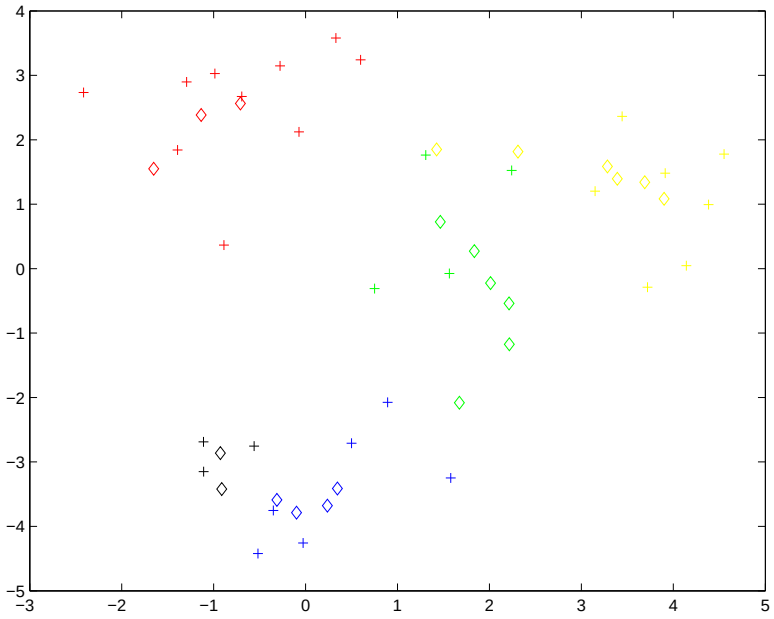
Sur chaque classe, nous avons mis à jour sa moyenne en la calculant avec tous les individus classés à l'intérieur. Puis nous avons réitérer le calcul des distances, le classement, etc... en gardant les moyennes successifs d'une même classe pour obtenir :



Nous avons ensuite chercher à discerner 4 classes dans notre échantillon :



puis 5 :



2 Interprétation

J'ai choisi de chercher plus de classe que de lois normales parce qu'on ne connait pas a priori le nombre de classe d'un échantillon.

On remarque que plus on cherche de classe, plus la variance de chaque classe est petite : les points d'une même classe sont très proches de leur moyenne respective.

On peut noter aussi qu'il n'y a pas besoin de beaucoup d'itérations pour trouver la moyenne d'une classe à partir d'une moyenne initiale quelconque.

3 Conclusion

L'algorithme de k-moyennes est capable de trouver des classes satisfaisantes la ou il n'y a pas de relation entre les individus de l'échantillon. Ainsi on peut distinguer des sous-classes dans une même classe d'un échantillon.

Pour avoir fait tourner l'algorithme plusieurs fois, nous n'avons pas toujours trouve un classement identique, d'ou l'importance de la moyenne initiale.

Ce TP était intéressant dans la mesure ou on se rend compte qu'on peut retrouver nos classes initiales assez rapidement : On peut donc en conclure que la méthode des k-moyennes converge assez rapidement.

Sur plusieurs dimensions, on pourrait distinguer des relations entre certaines dimensions, au sein d'une même classe.