A2DI - Comparaison des algorithmes k-means et clustering spectral

Quentin Baert - Antonin Carette

26 janvier 2016

Afin de comparer l'algorithme k-means à celui du clustering spectral, nous avons utilisé deux jeux de données :

- le jeu de données IRIS utilisés dans le TP précédent[1],
- des données représentant deux cercles concentriques[2].

L'algorithme spectral utilise deux paramètres σ et θ et les résultats de l'algorithme semble grandement dépendant de ces derniers. De plus il semble également que les valeurs des paramètres qui maximise ses résultats dépendent directement des données elles-mêmes. Pour trouver les valeurs des paramètres qui permettent d'obtenir les meilleurs résultats de l'algorithme spectral, nous avons implémenté une fonction `find_best_param()` (voir fichier `src/main.py`) qui, en fonction d'un jeu de données, exécute un nombre n de tests en faisant varié la valeur deux paramètres dans l'intervalle $[0.1, 1]$ par pas de 0.1 et renvoit les valeurs de σ et θ qui donne le moins d'erreurs en moyenne (sur les n runs).

Malgré ces dispositions, pour plusieurs exécution de la fonction `find_best_param()` sur le même jeu de données, la meilleur valeur des paramètres trouvés n'est pas toujours la même. L'algorithme spectral utilisant un k-means, nous supposons que les résultats de l'algorithme dépend également des clusters choisi par le k-means.

Les sections suivantes présentent les résultats obtenus avec les données présentées plus haut.

[1] chargé à l'aide de la fonction `sklearn.datasets.load_iris()`

[2] généré à l'aide de la fonction `sklearn.datasets.make_circles()`

Résultats sur les données IRIS

En travaillant sur les données IRIS, nous cherchons à séprarer 150 données en $k = 3$ clusters différents. Pour $n = 5$ runs, nous trouvons les résultats suivants ($\sigma = 0.4$, $\epsilon = 0.9$) :

- k-means : 104 erreurs,
- clustering spectral : 27 erreurs

On remarque ici que l'algorithme spectral est bien plus efficace que l'algorithme des k-means.

Résultats sur les cercles concentriques

En travaillant sur les données en forme de cercles concentriques, nous cherchons à séparer 100 points en $k = 2$ clusters différents. Pour $n = 5$ runs, nous trouvons les répartitions suivantes ($\sigma = 0.4$, $\theta = 0.2$):

On remarque encore une fois que le clustering spectral est plus efficace que les k-means. Nous nous interrogeons sur la régularité des erreurs commises par le clustering spectral sans y trouver d'explications. De plus, comme le montres les figures 3 et 4, l'introduction de bruit dans les données annule complètement l'efficacité du clustering spectral.

Conclusion

Le clustering spectral s'avère plus efficace que les k-means à condition de trouver les paramètres qui correspondent le mieux aux données que l'on souhaite clusteriser. On remarque également qu'un certain désordre dans les données peut annuler l'efficacité du clustering spectral.

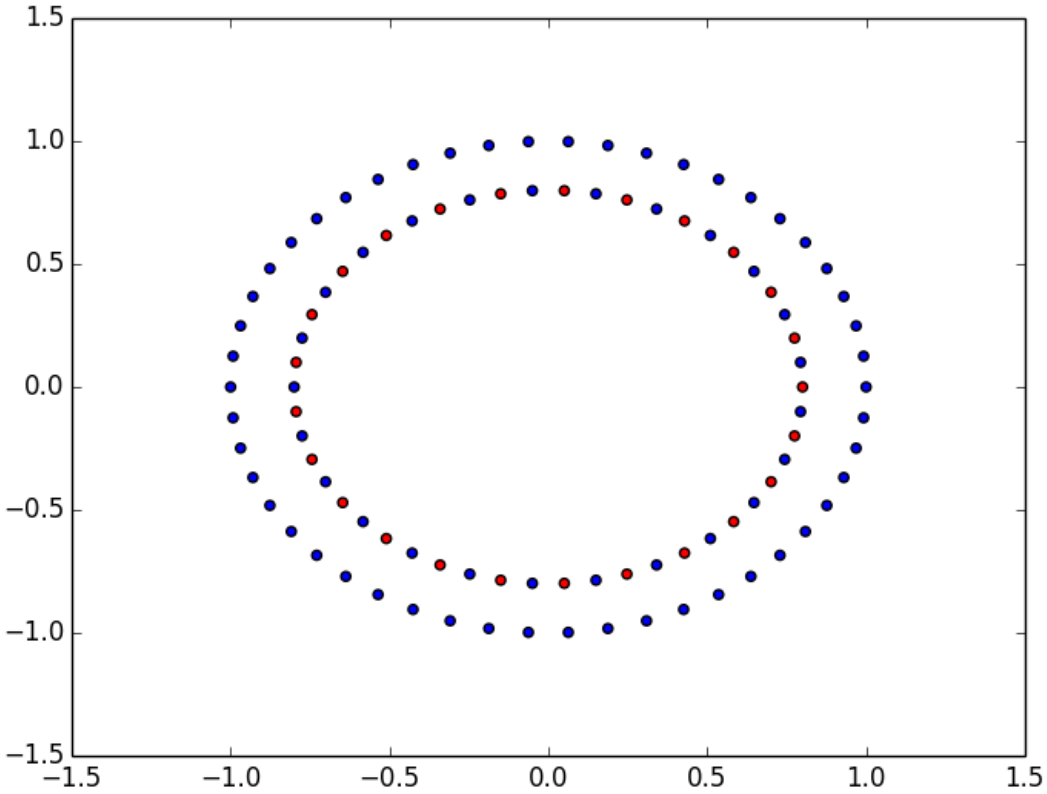

Figure 1: Résultat du clustering spectral sur les cercles concenriques

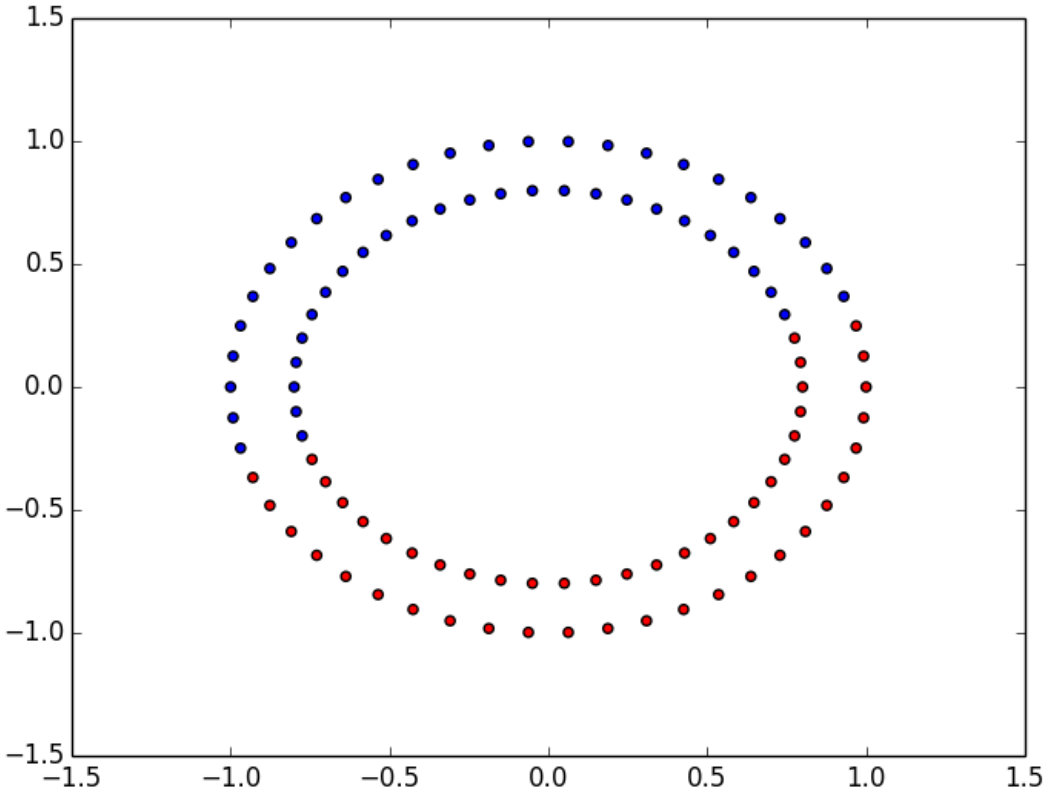

Figure 2: Résultat des k-means sur les cercles concenriques

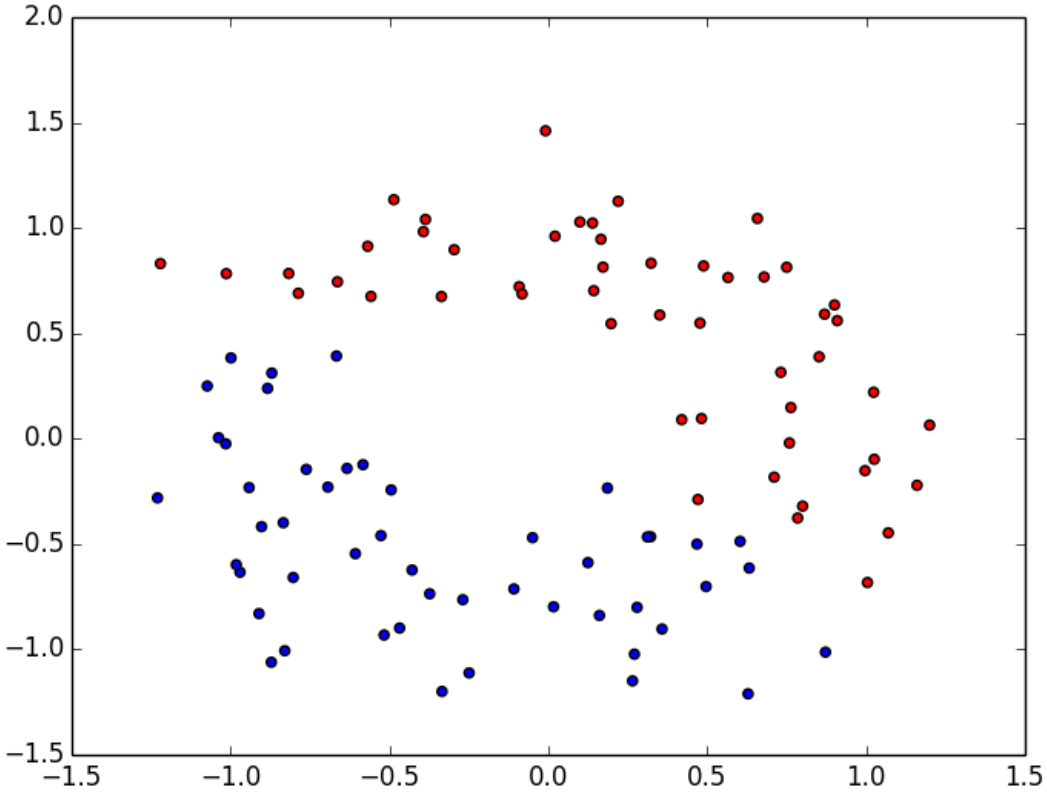

Figure 3: Résultat du clustering spectral sur les cercles concenriques bruités ($\sigma = 0.4$, $\theta = 0.3$)

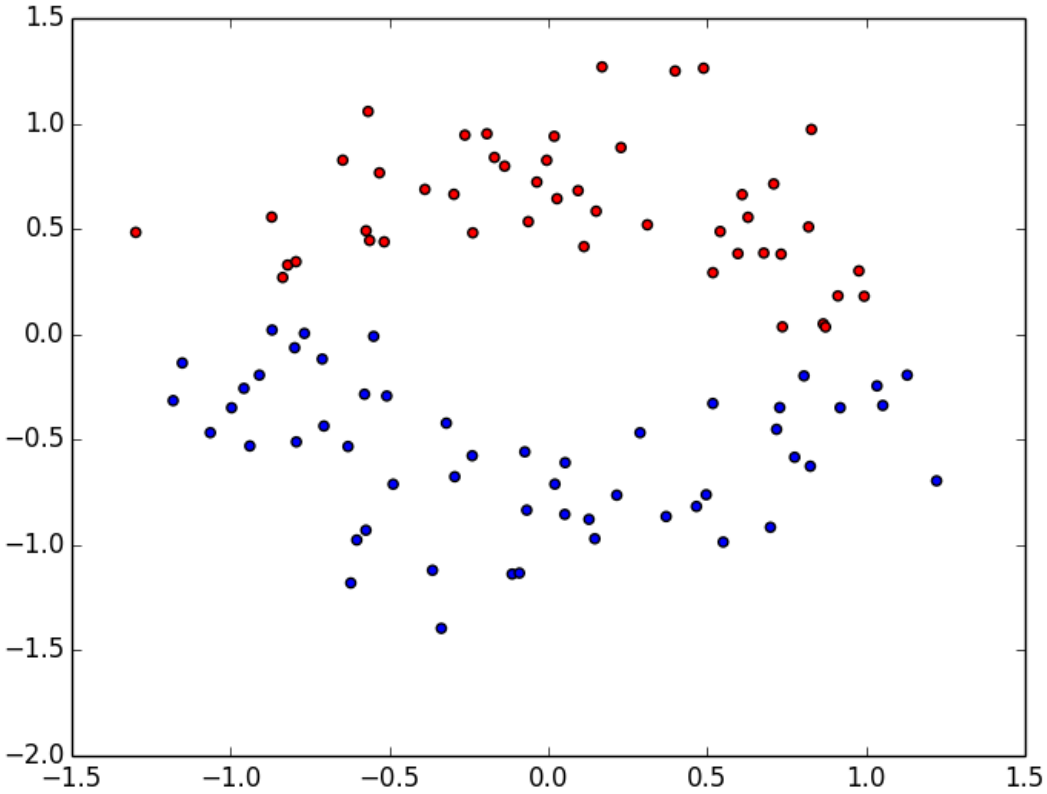

Figure 4: Résultat des k-means sur les cercles concenriques bruités