Master 1 SD

Nom : Coelho Fernandes
Prénom : Rodrigo César

Analyse et visualisation de données

Séance de TP 4
Density Based Clustering (DBSCAN)

1- Codage de DBSCAN

a. my_DBSCAN( )

```
def my_DBSCAN(X, eps, minpts, Visualisation=False):
    N, pp = np.shape(X)
    no_cluster = 0
    Dist = np.linalg.norm(X[:, np.newaxis] - X, axis=2)
    # eps = estime_EPS(Dist)
    # minpts = estime_MINPTS(X,Dist,eps)


    Visite = [False] * N
    y = -np.ones(N)  # tableau des labels des données, initialisé bruit (-1)
    Clusters = []

    for p in range(N):
        if not Visite[p]:
            Visite[p] = True
            Voisins = EpsilonVoisinage(p, X, Dist, eps)
            if len(Voisins) >= minpts:
                no_cluster += 1
                cluster = [p]
                y[p] = no_cluster
                cluster, y, Visite = etendre_cluster(X, y, Dist, cluster,
no_cluster, Voisins, Visite, eps, minpts)
                Clusters.append(cluster)

    return y
```

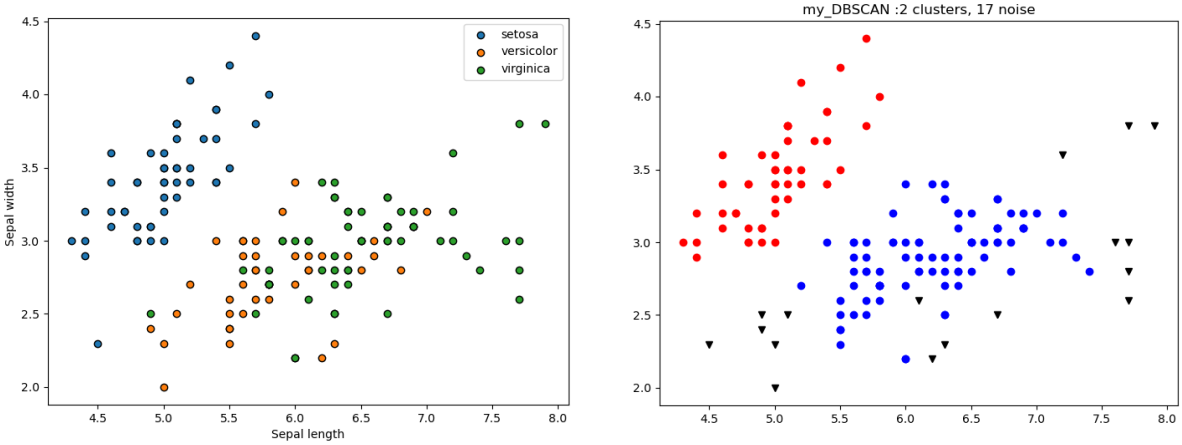
b. etendre_clusters( )

```
def etendre_cluster(X, y, Dist, Cluster, no_cluster, Voisins, Visite, eps,
minpts):
    for v in Voisins:
        if not Visite[v]:
            Visite[v] = True
            if y[v] == -1:
                y[v] = no_cluster
                Cluster.append(v)

            Voisins2 = EpsilonVoisinage(v, X, Dist, eps)
            if len(Voisins2) >= minpts:
                for vv in Voisins2:
                    if vv not in Voisins:
                        Voisins.append(vv)

    return Cluster, y, Visite
```

2- Analyse de DBSCAN

a. *Analyser le résultat de votre programme sur les données IRIS. Combien de Clusters sont déterminés pour les valeurs par défaut des deux hyper-paramètres?*

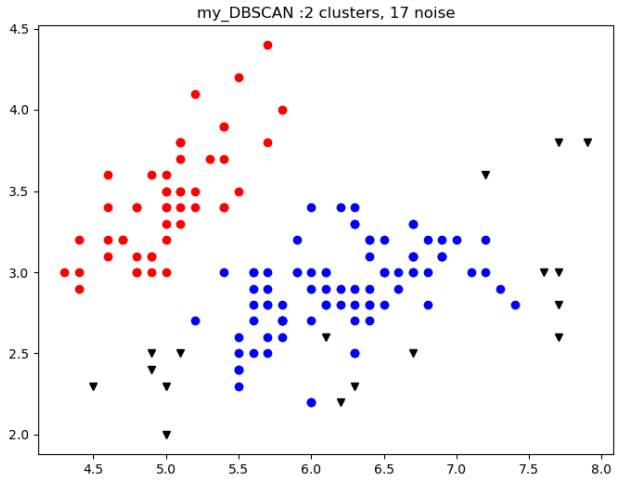
En utilisant les hyperparamètres par défaut (minPts=5 et eps=0,5), 2 clusters ont été détectés, ce qui signifie qu'aucune distinction n'a été trouvée entre les clusters Virginica et Versicolor, car ces clusters se chevauchent. Le cluster Setosa a été correctement identifié.



De plus, 17 points ont été classés comme du bruit, car les frontières des clusters sont des zones de faible densité et le nombre minimum de voisins pour la valeur de eps n'a pas été trouvé.

b. Comparer à la méthode DBSCAN de scikitlearn

La classification des points en clusters réalisée par scikit-learn présente le même nombre de clusters et de points indéterminés, et la distribution dans le plan à deux dimensions semble également être exactement la même. Cela indique que les deux méthodes sont en accord entre elles.

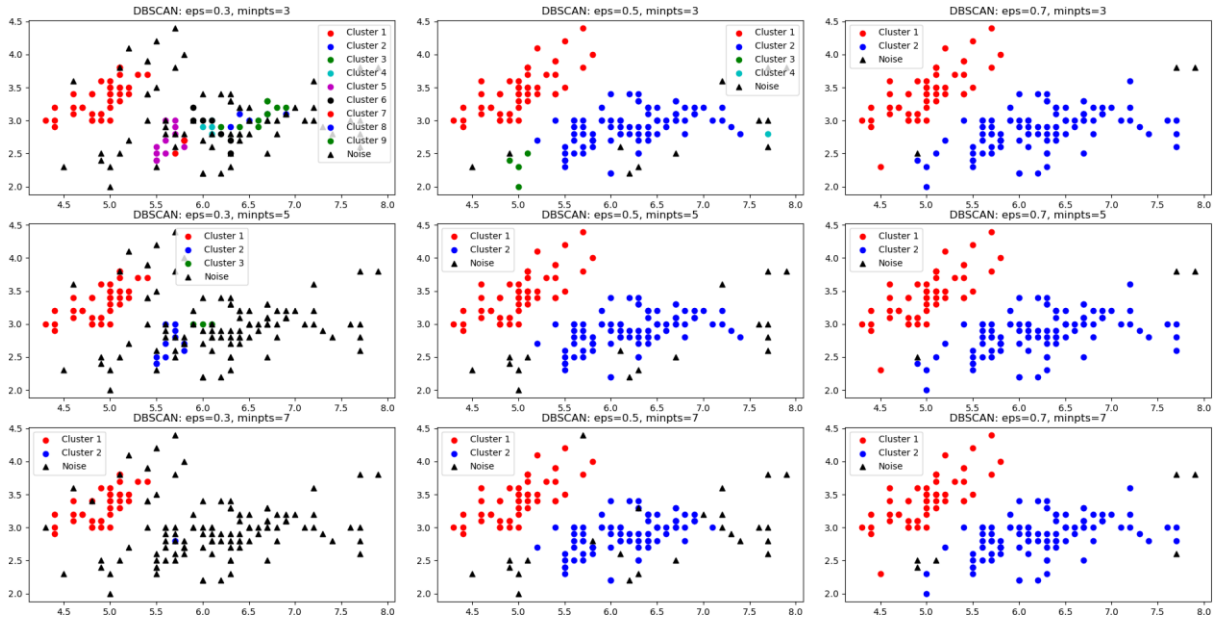


3- Influence des hyperparamètres eps et minpts

a) faire varier ces deux paramètres dans les plages suivantes et noter le nombre de clusters trouvés pour chaque configuration ep in [0.3,0.5,0.7] minpts $\in [3, 5, 7]$

Bruit		eps		
		0,3	0,5	0,7
minpts	3	71	12	3
	5	99	17	3
	7	114	25	6

K		eps		
		0,3	0,5	0,7
minpts	3	9	4	2
	5	3	2	2
	7	2	2	2



b) quel est le paramètre le plus influent ?

Parmi les deux hyperparamètres, eps est celui qui a l'influence prédominante sur la classification des points. Les résultats montrent que plus l'hyperparamètre eps est grand, moins il y a de bruit et moins il y a de clusters. Cela se produit parce qu'un rayon plus grand facilite la satisfaction de la condition minPts, donc l'algorithme continue plus longtemps dans le processus de classification d'un même cluster et parvient à classifier même les points aux extrémités, où la densité est faible.

En revanche, l'augmentation du paramètre minPts a une forte influence sur la quantité de bruit, car toutes les régions n'ont pas une densité suffisamment élevée pour satisfaire des valeurs élevées de minPts. La quantité de clusters diminue avec l'augmentation de minPts, mais de manière moins marquée, car le paramètre eps a une influence plus prédominante. Cependant, une valeur plus élevée de minPts inhibe la propagation des clusters et la création de nouveaux clusters dans des zones isolées de plus grande densité.

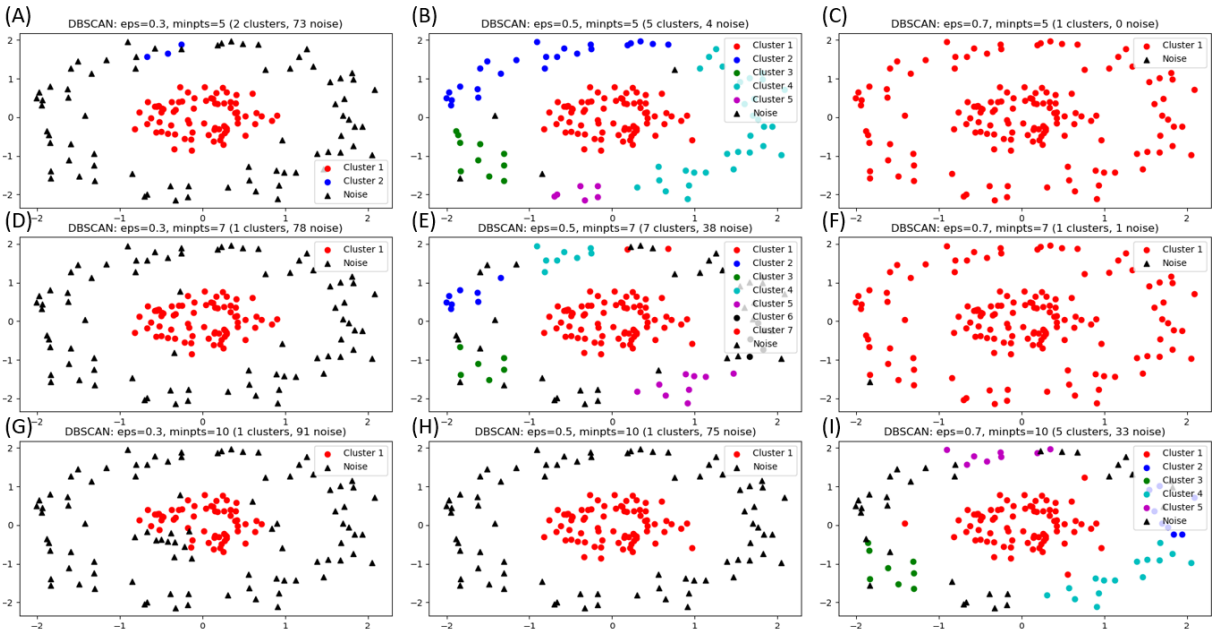
4- Analyse des données circulaires

a. faire varier les deux hyperparamètres dans les plages suivantes et noter le nombre de clusters trouvés pour chaque configuration

$ep \in [0.3, 0.5, 0.7]$ $minpts \in [5, 7, 10]$

Bruit		eps		
		0,3	0,5	0,7
minpts	5	73	4	0
	7	78	38	1
	10	91	75	33

K		eps		
		0,3	0,5	0,7
minpts	5	2	5	0
	7	1	7	1
	10	1	1	5



b. quel est le paramètre le plus influent ?

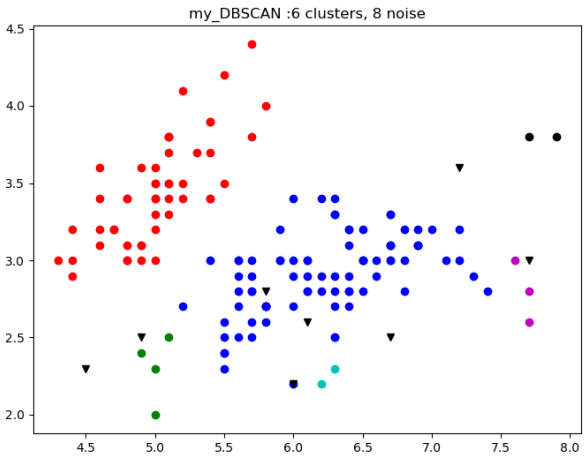
Pour cette distribution de données, on observe que le cercle central possède une densité élevée et, en conséquence, il obtient une classification satisfaisante dans tous les cas. En revanche, le cercle extérieur présente une densité plus faible, ce qui entraîne des problèmes de classification lorsque les valeurs des hyperparamètres ne sont pas bien définies, avec eps continuant à avoir une influence prédominante.

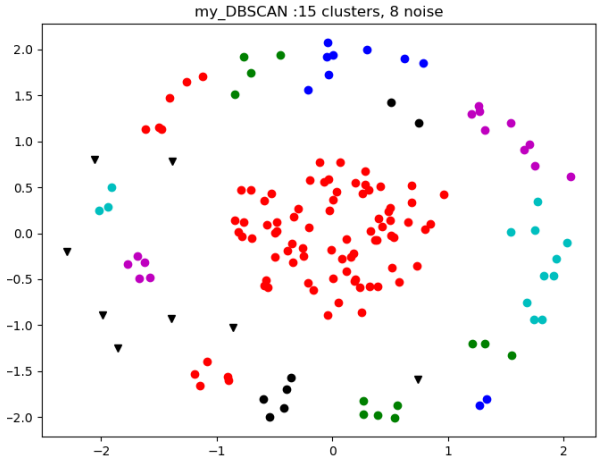
Pour les combinaisons d'hyperparamètres (A, D, G, H), le rayon n'est pas suffisant pour englober un nombre satisfaisant de voisins, et le cercle présente une grande zone d'indécision. Dans les résultats de B, E et I, le cercle extérieur a été divisé en plusieurs clusters qui reflètent les différentes interruptions du parcours de l'algorithme, causées par les zones de faible densité. Enfin, C et F montrent qu'un grand rayon et une faible condition minimale ont permis à l'algorithme de passer d'un cercle à l'autre, créant ainsi un seul cluster.

5- Programmer les méthodes d'estimation des deux hyperparamètres, en respectant les interfaces ci-dessous.

```
def estime_EPS(Dist):
    # estimation du rayon du epsilon voisinage
    N = Dist.shape[0]
    Diag = np.eye(N)*1000
    EPS = np.percentile(np.min(Dist+Diag,axis=0),95)

    return EPS

def estime_MINPTS(X,Dist,eps):
    # estimation de minpts dans le epsilon voisinage
    NVoisins = []
    N,pp =np.shape(X)
    for p in range(N):
        NVoisins = NVoisins+[len(EpsilonVoisinage(p,X,Dist,eps))]
        MINPTS = math.ceil(np.percentile(np.asarray(NVoi-
sins,dtype=np.float64),5))

    return MINPTS
```

6- Tester les deux estimateurs les deux jeux de données. Combiens de clusters sont obtenus ? Que peut-on conclure ?





Les estimateurs des hyperparamètres éliminent les outliers de la distance entre les points et la quantité de voisins afin de définir une valeur qui puisse mieux représenter les données. Pour le jeu de données Iris, eps=0,475 et minPts=2 ont été calculés, tandis que pour les cercles générés, nous avons eps=0,333 et minPts=2.

Les deux combinaisons de paramètres se sont révélées efficaces pour réduire les zones d'indécision, et très peu de points ont été classés comme du bruit. Cependant, la définition des clusters n'a été satisfaisante dans aucun des cas. Dans le jeu de données Iris, des zones de faible densité à la périphérie des données ont été classées comme de petits clusters distincts. Nous pouvons affirmer que DBSCAN n'est pas une bonne technique pour séparer correctement les clusters du jeu de données Iris, indépendamment des hyperparamètres, en raison du chevauchement entre Virginica et Versicolor.

Dans l'autre jeu de données, la grande différence de densité entre les deux cercles rend impossible l'existence d'une seule valeur d'eps et de minPts idéale pour les deux ensembles. Les paramètres estimés ne conviennent pas bien au cercle extérieur, ce qui entraîne sa division en plusieurs clusters.

7- Quels sont les problèmes non résolus par DBSCAN?

Le succès de l'algorithme dépend fortement de la bonne sélection des paramètres eps et minPts. Si ces paramètres ne sont pas choisis correctement, DBSCAN peut générer des clusters trop petits (comme on le voit dans les images B, E et I), ou au contraire

fusionner des clusters qui ne devraient pas être fusionnés (comme c'est le cas dans C et F).

De plus, DBSCAN fonctionne bien lorsque les clusters sont séparés par des régions de faible densité, mais il échoue lorsque les clusters ont des densités très différentes, comme les deux cercles générés par la fonction make_circles(). L'algorithme ne peut pas s'adapter à plusieurs densités, car les paramètres eps et minPts sont fixes.

Enfin, DBSCAN est incapable de différencier deux clusters qui se chevauchent, comme Virginica et Versicolor dans le jeu de données Iris, car il n'y a pas de différence de densité entre ces deux clusters pour créer une interruption dans le cheminement de l'algorithme.

Un autre exemple d'échec dans la détermination des clusters est l'ensemble de données sur le cancer du sein, où même les hyperparamètres estimés (eps=85.35 et minpts=2) ne sont pas capables de distinguer les clusters qui se mélangent.