Réalisé par : Mbaye THIAM Encadré par : Pr. LUCIANI Xavier

Analyse de Données, Machine Learning :

SVM (Support Vecteur Machine) linéaires et non linéaires

INTRODUCTION :

L'objectif de ce projet est, d'explorer les mécanismes fondamentaux des SVM, et de comprendre comment ces algorithmes peuvent être conçus et appliqués pour résoudre des problèmes concrets de machine Learning.

Les SVM sont particulièrement efficaces pour résoudre des problèmes de classification, même dans des contextes où les données ne sont pas linéairement séparables. En utilisant des concepts tels que le problème primal et dual, les SVM offrent une flexibilité remarquable dans la résolution de diverses tâches d'apprentissage.

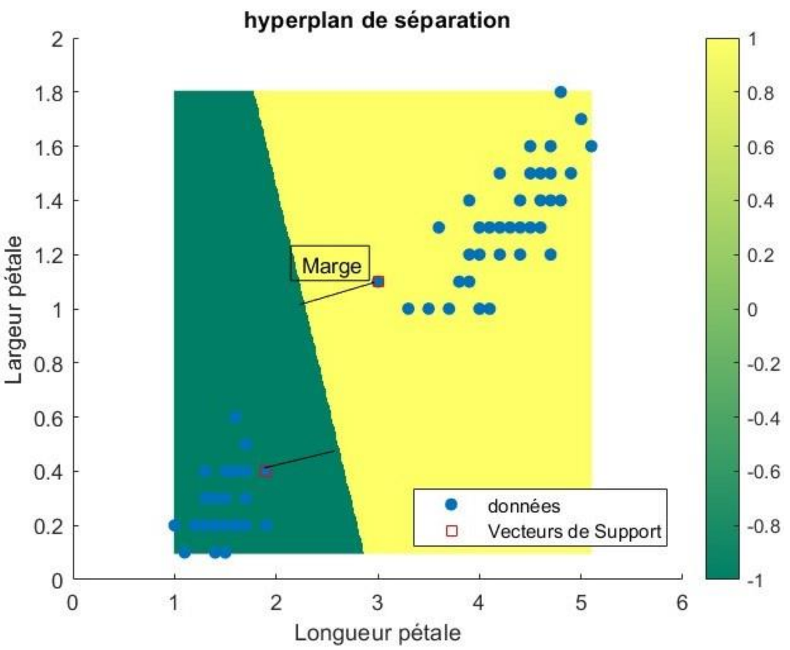
A. **Cas linéairement séparable « Résolution du problème primal »**

On cherche à trouver les paramètres optimaux d'un hyperplan de séparation linéaire, en optimisant la marge et la précision de la classification.

On réécrit le problème primal comme un problème de programmation quadratique :

$$(w_0,\ w^*) = argmin\left\{\frac{1}{2} * ||W||^2\right\} \quad ||W||^2 = W^T I_N W$$

$$\text{Sous les contraintes : qqs m, } y_m\left(w_0 + w^T x_m\right) \geq 1 \quad :$$



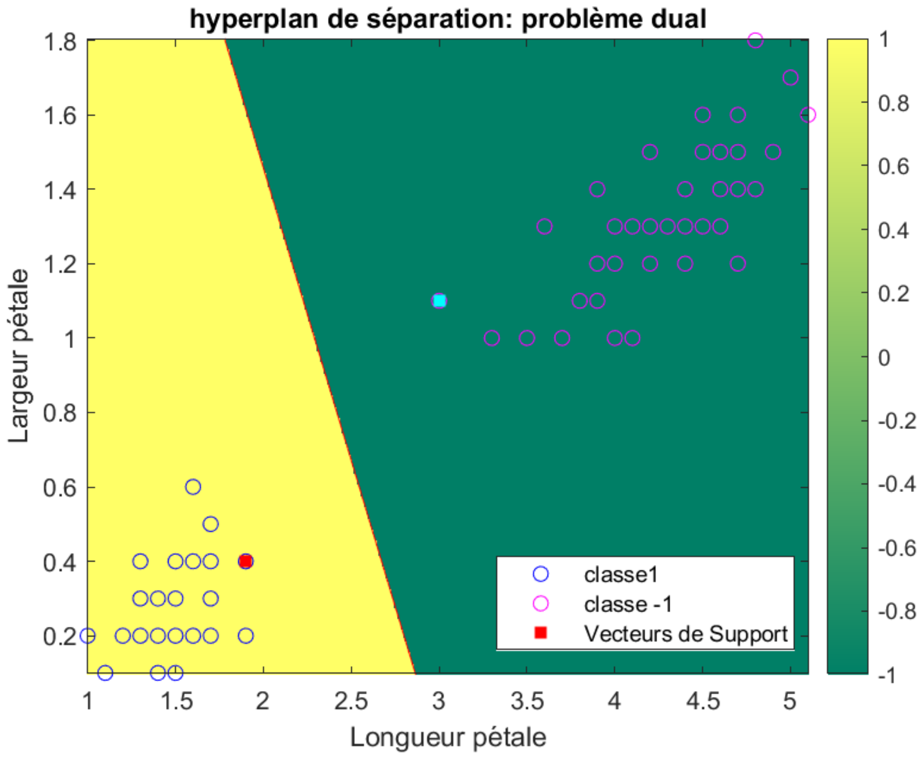
La figure illustre les résultats de l'algorithme des SVM pour résoudre le problème primal dans le contexte de données Iris. L'hyperplan de séparation, démontre une classification précise des 2 différentes classes d'Iris avec une marge significative. La précision de classification atteint 100%, soulignant l'efficacité du modèle dans la séparation des classes. Ces résultats confirment la capacité des SVM à réaliser des classifications linéaires avec une grande précision.

B. Cas linéairement séparable « Résolution du problème dual »

L'objectif reste le même, trouver les paramètres optimaux d'un hyperplan de séparation linéaire, mais en introduisant de nouvelles variables à optimiser appelées multiplicateurs de Lagrange :

$$\alpha^* = argmax\left\{ \sum_{m=1}^{M} \alpha_m - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} \alpha_i \alpha_j y_i y_j x_i^T x_j \right\} \text{ Sous } \textit{les contraintes qqs} \text{ m, } \left\{ \alpha_m > 0 \ \ et \sum_{m=1}^{M} \alpha_m y_m = 0 \right\}$$

$$w^* = \sum_{m=1}^{M} \alpha_m y_m x_m \quad \text{et} \quad y_m\left(w_0^* + w^{*T} x_m\right) = 1 \ \ \textbf{pour} \ \alpha_m \neq 0$$



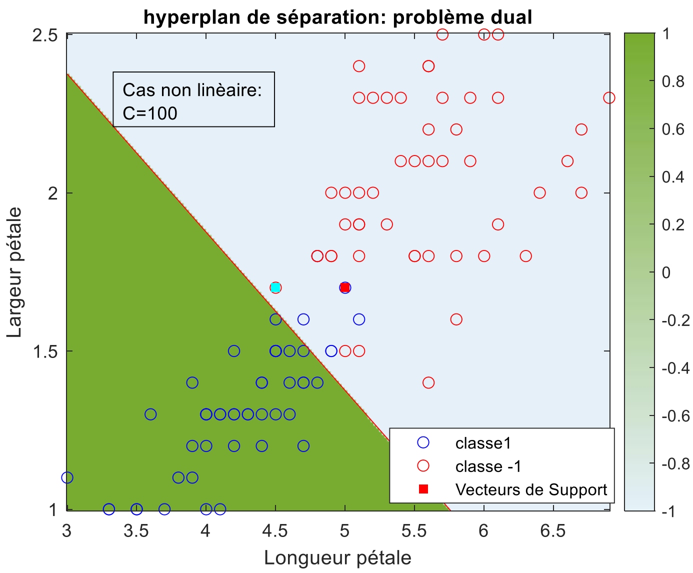
On voit que le cas dual sous contraintes optimales est aussi efficace avec une précision de classification de 100% et avec une marge importante.

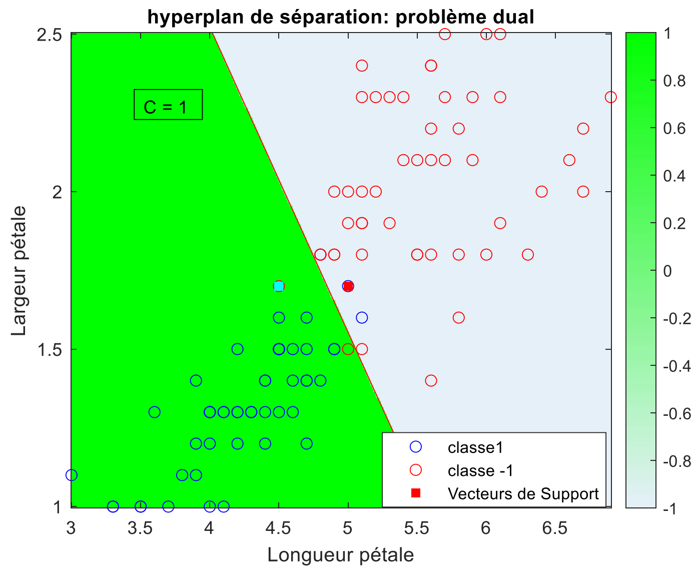
Ces résultats confirment la pertinence du choix de résoudre le problème dual dans le contexte du SVM linéairement séparable.

C. Cas non linéairement séparable, résolution du problème dual

Dans cette section, nous explorons le cas des données non linéairement séparables. Confrontés à un mélange complexe des classes, l'utilisation d'un séparateur linéaire ne serait pas optimale. Nous allons résoudre ce défi en utilisant une approche basée sur le problème dual des SVM. Cette stratégie nous permettra de généraliser la séparation des classes au-delà des frontières linéaires, améliorant ainsi la capacité du modèle à traiter des données plus complexes et non linéaires.

NB : On change les conditions sur les contraintes



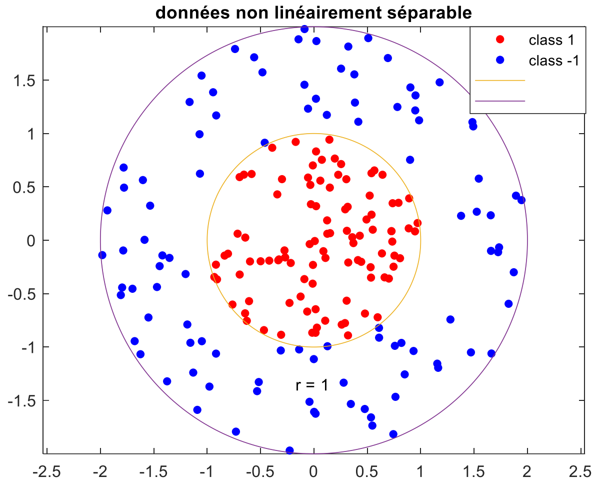


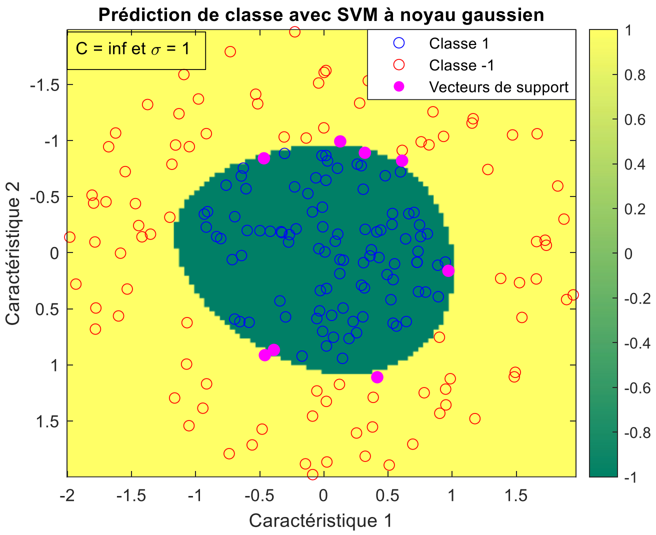
Les deux figures illustrent la classification des données avec un modèle linéaire pour 2 valeurs du paramètre de régularisation (C=100 et C=1). Malgré une séparation linéaire, les résultats montrent que le modèle n'atteint pas une précision de 100%, ce qui indique que les données ne sont pas parfaitement linéairement séparables. La variation de C influence la performance, on peut voir que C petit donne de meilleurs résultats mais induit plus d'erreurs d'apprentissage.

Donc pour optimiser la classification sans contrepartie il est nécessaire d'explorer d'autres approches, notamment les noyaux non linéaires.

D. Exemple de classifieur non linéaire : noyau gaussienne

Dans cette partie, nous allons explorer un classifieur non linéaire a noyau gaussien sur les données de la figure ci-dessous. Dans le cadre du noyau on adapte le calcul des paramètres de l'hyperplan de séparation avec la fonction du noyau (cf. cours 4)





La figure ci-dessus illustre la performance de notre classifieur non linéaire à noyau gaussien sur une représentation binaire. La séparation non linéaire capture efficacement les structures complexes présentes dans les données avec une précision de 100%.

NB. Ces performances sont calculées sur la base de l'entrainement

```
Accuracy: 100.00%
Precision: 1.00
Matrice de Confusion :
    100      0
      0    100
```

E. Classifieur polynomial pour le problème des iris

Le classifieur polynomial offre une séparation assez simple des données, bien que
quelques individus ne soient pas correctement classés.
Il serait peut-être nécessaire d'explorer d'autres types de noyaux ou d'ajuster les paramètres du modèle pour améliorer la performance. En revanche, le classifieur noyau polynomial (degré 3) avec C=1 offre une meilleure précision de classification.

```
Accuracy: 95.00%
Precision: 0.96
Matrice de Confusion :
     47      3
      2     48
```