

# TP N°3 : Support Vector Machine (SVM)

Maxence Lamure

maxence.lamure@etu.umontpellier.fr

M2 SSD-MIND

27/09/2024

## 1 Introduction aux SVM

Les Support Vector Machines (SVM) représentent une famille d’algorithmes d’apprentissage supervisé, largement utilisés pour les problèmes de classification et, dans une moindre mesure, de régression. Le principe des SVM repose sur la recherche d’un hyperplan séparateur optimal entre différentes classes de données. Cet hyperplan est défini de manière à maximiser la marge entre les points de chaque classe les plus proches de cette frontière, appelés vecteurs de support. La fonction de classification pour un SVM est donnée par :

$$f(x) = \text{sign} \left( \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right)$$

où  $\alpha_i$  sont les coefficients du modèle,  $y_i$  les labels,  $K(x_i, x)$  est le noyau choisi, et  $b$  est le biais.

Dans le cas où les données sont linéairement séparables, le SVM détermine un hyperplan qui maximise la distance entre les vecteurs de support des différentes classes. Cependant, toutes les données ne sont pas toujours linéairement séparables. Pour répondre à cette problématique, les SVM utilisent la méthode du noyau (kernel) qui permet de transformer les données en un espace de dimension supérieure où une séparation linéaire devient possible.

Dans ce TP, nous allons principalement travailler avec deux types de noyaux :

- Cas linéaire, adapté lorsque les données peuvent être séparées par un hyperplan.
- Cas polynomial, utile pour capturer des relations non linéaires en augmentant la complexité de la frontière.

L’objectif de ce travail est de comparer les performances des SVM en fonction des noyaux utilisés et de l’ajustement de leurs hyperparamètres. En modifiant des paramètres

comme le paramètre de régularisation ou les paramètres des noyaux, nous chercherons à comprendre comment ces manipulations influencent la qualité de la classification.

## 2 Application

### 2.1 Q1 : Noyau linéaire

À partir du code donnée en annexe, nous classifions la classe 1 contre la classe 2 en utilisant uniquement les deux premières variables caractéristiques de l'ensemble Iris et un noyau linéaire. Nous séparons équitablement les données d'entraînement et de test en deux échantillons distincts. Dans le cas d'un noyau linéaire, la fonction de noyau est définie comme suit :

$$K(x, x') = \langle x, x' \rangle$$

L'analyse des résultats montre que le modèle SVM avec un noyau linéaire atteint un score de 72% sur les données d'entraînement, mais celui-ci chute légèrement à 68% sur les données de test, ce qui suggère un problème de généralisation. Cela pourrait indiquer que le modèle est légèrement surentraîné ou qu'il ne capture pas toute la complexité des données.

### 2.2 Q2 : Noyau polynomial

Nous utilisons un noyau polynomial pour comparer ses performances au noyau linéaire. Le noyau polynomial, permettant de capturer des relations non linéaires, est donné par :

$$K(x, x') = (\alpha + \beta \langle x, x' \rangle)^\delta, \quad \delta > 0$$

où  $\alpha$ ,  $\beta$ , et  $\delta$  sont des paramètres ajustables.

Voici les paramètres optimaux trouvés avec la validation croisée :

```
{ 'C': 0.03162277660168379, 'degree': 1, 'gamma': 10.0, 'kernel': 'poly' }
```

Le modèle SVM avec un noyau polynomial atteint un score de 74% sur les données d'entraînement et de 72% sur les données de test, ce qui suggère une efficacité sensiblement équivalente entre ce modèle et celui avec un noyau linéaire.

La Figure 1 met en évidence les frontières de décision obtenues avec un noyau linéaire et un noyau polynomial, ainsi que la distribution des observations selon ces deux modèles.

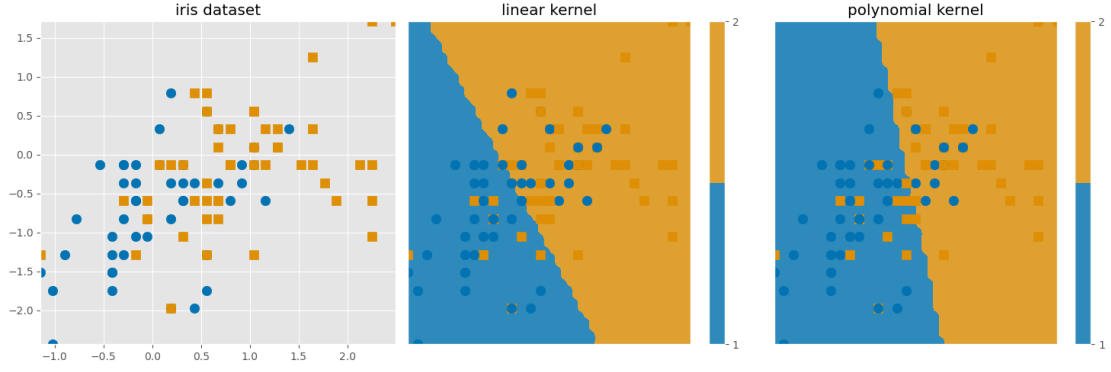


Figure 1: Performances des SVM appliqués au jeu de données iris pour un noyau linéaire et polynomial

### 2.3 Q3 bonus : SVM GUI

En diminuant le paramètre  $C$  dans un SVM avec un noyau linéaire, on observe que le modèle devient plus tolérant aux erreurs sur les données d'entraînement. Concrètement, cela se traduit par une marge plus large entre les classes, au prix de certaines mauvaises classifications. Le modèle privilégie ainsi une meilleure généralisation, en évitant de se concentrer trop strictement sur chaque point d'entraînement. Ce compromis est représenté par le problème d'optimisation suivant :

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

sujet aux contraintes  $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$  et  $\xi_i \geq 0$ .

Dans le cadre d'un jeu de données déséquilibré, comme celui utilisé ici (90% des points dans une classe et 10% dans l'autre), une valeur plus faible de  $C$  tend à favoriser la classe majoritaire. En effet, la frontière de décision devient moins précise et plus influencée par la majorité des points, ce qui peut entraîner une dégradation de la performance sur la classe minoritaire, avec davantage de points mal classifiés dans cette catégorie.

## 3 Classification de visages

Pour classer les visages, nous utilisons l'ensemble de données LFW (Labeled Faces in the Wild) puis nous sélectionnons deux personnes spécifiques (Tony Blair et Colin Powell) dont nous disposons au moins 70 photos et réduisons la taille des images par optimisation d'espace. Nous préparons ensuite les données dont un échantillon se trouve dans la Figure 2 pour une tâche de classification binaire.



Figure 2: Echantillon des données LFW

### 3.1 Q4 : Paramètre de régularisation

Nous étudions l'influence du paramètre de régularisation  $C$  en ajustant les valeurs sur une échelle logarithmique entre  $10^{-5}$  et  $10^5$  et en observant les scores :

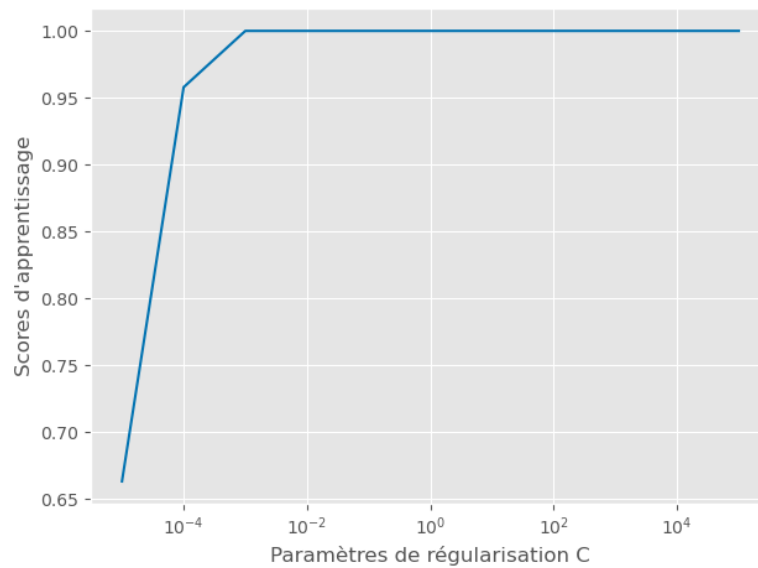


Figure 3: Graphe montrant le score d'apprentissage en fonction du paramètre de régularisation  $C$

Comme montré dans la Figure 3, le meilleur paramètre de régularisation est  $C = 10^{-3}$ , pour un score de 1. Ce score reste le même pour un  $C$  plus grand, ce qui montre la stabilité de la classification.

Nous réentraînons à présent le modèle avec ce  $C$  optimal et réalisons les prédictions finales sur les données de test. Nous comparons ensuite la précision obtenue à celle attendue d'un modèle prédictif aléatoire, servant de référence pour évaluer les performances.

Nous obtenons alors une précision de 91% et un niveau de chance de 62%. Le modèle affiche donc une performance nettement meilleure que l'aléatoire, ce qui montre qu'il a bien appris à différencier les classes à partir des données fournies.

Les résultats sont illustrés dans la Figure 4.



Figure 4: Echantillon de prédiction du modèle pour les données LFW

La Figure 5 présente quant à elle les coefficients du modèle linéaire appris sous forme de carte de chaleur sur une grille bidimensionnelle :

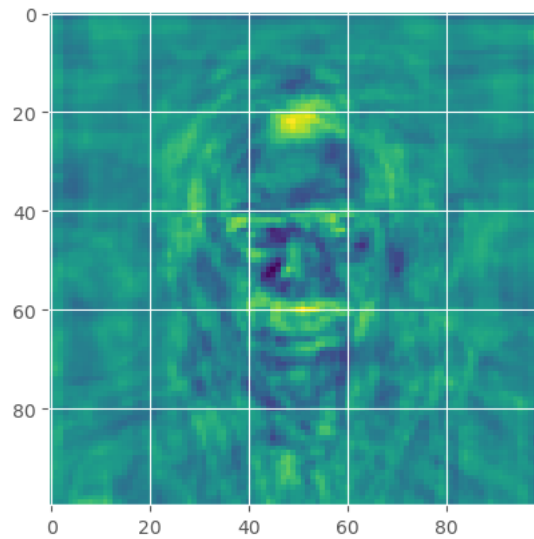


Figure 5: Coefficients du modèle linéaire appris sous forme de carte de chaleur

Les variations de couleurs représentent les différences d'intensité des coefficients ou des données dans cette grille. Typiquement, des zones plus claires indiquent des valeurs plus élevées, tandis que des zones plus sombres signalent des valeurs plus faibles.

Nous constatons alors que le classifieur se base principalement sur certains éléments tels que l'implantation des cheveux au-dessus du crâne et sur les côtés, les sourcils/yeux, le nez et la bouche pour sa prédiction.

### 3.2 Q5 : Variables de nuisance

Nous ajoutons ici des variables de nuisance pour évaluer l'impact sur les performances du modèle.

Sans variables de nuisance, le modèle atteint un score de 1 pour les données d'entraînement et de 93% pour les données de test, montrant une très bonne performance et sa capacité à généraliser sur de nouvelles données.

Avec variables de nuisance en revanche, le modèle atteint certes un score de 1 pour les données d'entraînement mais chute significativement à 47% pour les données de test (inférieur au niveau de chance à 62%), indiquant une mauvaise robustesse face à de nouvelles données malgré une bonne performance pour les données d'entraînement.

### 3.3 Q6 : Réduction de dimensions ACP

Après avoir analysé l'impact des variables de nuisance sur les performances du modèle, il peut être utile d'étudier une autre technique de prétraitement des données : la réduction de dimensions. Grâce à l'Analyse en Composantes Principales (ACP), l'objectif

est de diminuer le nombre de variables tout en conservant un maximum d'information. Cette approche peut améliorer la performance du modèle en éliminant le bruit et en simplifiant la structure des données. Nous appliquerons donc l'ACP sur les données bruitées pour évaluer son effet sur la précision du modèle SVM, en testant trois scénarios : avec 40, 80 et 120 composantes.

Ainsi, après un temps de calcul relativement long, on obtient les résultats suivants :

- 40 composantes : Generalization score for linear kernel: 0.7473684210526316 (train), 0.5684210526315789 (test)
- 80 composantes : Generalization score for linear kernel: 0.868421052631579 (train), 0.4789473684210526 (test)
- 120 composantes : Generalization score for linear kernel: 0.8631578947368421 (train), 0.5421052631578948 (test)

On observe une tendance au sur-apprentissage à mesure que le nombre de composantes augmente. En effet, avec 80 et 120 composantes, le score d'entraînement reste élevé, mais le score de test diminue. Cela suggère que le modèle devient trop complexe et s'adapte trop étroitement aux données d'entraînement, au détriment de la généralisation sur des données non vues.

La réduction du nombre de composantes à 40 permet d'obtenir un meilleur équilibre entre les scores d'entraînement et de test, bien que les scores globaux soient plus faibles que dans les cas avec plus de composantes. Cela montre que l'ACP aide à réduire le bruit et à améliorer la généralisation jusqu'à un certain point. Au-delà d'un certain nombre de composantes, l'ajout de plus de dimensions semble réintroduire du bruit ou rendre le modèle plus sujet au sur-apprentissage.