TP3 : CLASSIFICATION

SVM à noyau RBF

Dans ce TP, vous allez travailler sur un jeu de données médical réel utilisé pour diagnostiquer des tumeurs du sein. Après une réduction en 2 dimensions, les données présentent une structure non linéaire difficile à séparer avec un modèle linéaire.

Votre objectif est de comprendre comment le SVM à noyau RBF peut capturer cette structure complexe et produire un modèle performant.

PARTIE 1 — Télécharger et comprendre le dataset

Objectif : Découvrir un dataset réel et sa structure en 2 dimensions.

1. Télécharger le fichier : breast_cancer_svm_rbf_2d_ready.csv

2. Afficher les premières lignes du dataset.

3. Identifier ce que représentent les colonnes : feature1, feature2, target.

4. Quelles difficultés observe-t-on lorsqu'on projette des données complexes en 2D ?

PARTIE 2 — Visualiser les données en 2D

Objectif : Observer la distribution des classes et déterminer si une séparation linéaire est possible.

1. Afficher un nuage de points avec feature1 en X et feature2 en Y.

2. Colorer les points selon la classe cible (target).

3. Interpréter la répartition des classes.

4. Pourquoi un modèle linéaire pourrait-il être insuffisant ici ?

PARTIE 3 — Séparer les données en entraînement et test

Objectif : Préparer une évaluation correcte du modèle.

1. Diviser les données en 80% entraînement et 20% test.

2. Vérifier les tailles des deux ensembles.

3. Pourquoi choisir un random_state fixe ?

PARTIE 4 — Entraîner un SVM linéaire (modèle de référence)

Objectif : Montrer les limites d'une séparation linéaire.

1. Créer un modèle SVM linéaire.

2. L'entraîner sur les données d'entraînement.

3. Mesurer son accuracy sur le test.

4. Qu'indique cette performance sur la nature des données ?

PARTIE 5 — Entraîner un SVM RBF

Objectif : Utiliser un modèle capable de capturer une séparation non linéaire.

1. Créer un modèle SVM à noyau RBF (C=1, gamma=1).

2. L'entraîner sur les données.

3. Générer la Matrice de confusion.

4. Évaluer l'accuracy,Recall,Precision du modèle.

5. Analyser les métriques d'évaluation

5. Pourquoi le noyau RBF s'adapte mieux aux données ?

PARTIE 7 — Étude du paramètre gamma (γ)

Objectif : Comprendre l'effet de gamma sur la complexité de la frontière.

1. Tester plusieurs valeurs de gamma (0.01, 0.1, 1, 10).

2. Pour chaque gamma, mesurer accuracy train et test.

3. Que se passe-t-il si gamma est très petit ?

4. Que se passe-t-il si gamma est très grand ?

PARTIE 8 — Étude du paramètre C

Objectif : Comprendre le compromis entre marge large et penalisation des erreurs.

1. Tester plusieurs valeurs de C (0.1, 1, 10, 100).

2. Pour chaque C, mesurer accuracy train et test.

3. Quand C est petit, le modèle accepte-t-il plus ou moins d'erreurs ?

4. Quand C est grand, la marge devient-elle large ou stricte ?

5. Quel C semble un bon compromis ?

PARTIE 9 — Matrice de confusion, Accuracy, Precision, Recall

Objectif : Interpréter les métriques de classification en contexte médical.

1. Choisir la meilleure combinaison C/gamma.

2. Calculer accuracy, precision et recall.

3.Générer la Matrice de confusion

4. Quelle métrique est la plus critique pour la classe maligne ?

5. Que signifie un recall faible pour la classe maligne ?

6. Pourquoi l'accuracy seule peut-elle être trompeuse ?