Option : Ml OIVM
Module : TP ATDN

TP4 : Classification binaire avec SVM, Mélange de modèles, et Modèles Probabilistes Mixtes

Ce TP a pour but de :

- ✓ Explorer les SVM pour la classification binaire.
- ✓ Expérimenter avec des modèles combinés (mélanges de modèles).
- ✓ Appliquer des modèles probabilistes mixtes comme les *Gaussian Mixture Models (GMM)* pour la classification.
- ✓ Comparer les performances des différents modèles.
- ✓ Utiliser des outils avancés pour l'évaluation et l'optimisation des modèles.

Exercice 1 : Classification avec SVM

Comprendre le fonctionnement des SVM et leur application sur un dataset comme *SMS Spam Collection*.

1. Chargez et prétraitez le dataset SMS Spam Collection (vectorisation des messages et encodage des labels).
2. Divisez les données en ensembles d'entraînement et de test (70/30). Pourquoi la stratification est-elle importante ici ?
3. Entraînez un modèle SVM avec un noyau linéaire et affichez les performances (précision, rappel, F1-score).
4. Tracez la matrice de confusion et analysez les résultats. Quelles erreurs sont les plus fréquentes ?
5. Générez et interprétez la courbe ROC-AUC pour ce modèle.

Exercice 2 : Mélange de modèles (Voting Classifier)

Combiner plusieurs modèles pour améliorer la robustesse de la classification.

1. Créez et entraînez trois modèles de base : Naïve Bayes, Régression Logistique et SVM (noyau linéaire). Comparez leurs performances respectives sur le dataset.
2. Combinez ces modèles en utilisant un *Voting Classifier* avec un vote *hard* (majorité). Évaluez les performances.
3. Répétez avec un vote *soft* (moyenne des probabilités). Quel type de vote est plus performant dans ce cas ?
4. Tracez la courbe ROC-AUC pour les modèles individuels et le Voting Classifier. Analysez les résultats.
5. Pourquoi le mélange de modèles peut-il surpasser les performances des modèles individuels ?

Exercice 3 : Classification avec Gaussian Mixture Models (GMM)

Appliquer des modèles probabilistes mixtes pour la classification.

1. Expliquez le principe des *Gaussian Mixture Models* (GMM) et leur rôle dans la classification.
2. Implémentez un modèle GMM sur les données vectorisées (2 classes). Entraînez-le sur l'ensemble d'entraînement.
3. Évaluez les performances du modèle (précision, rappel, F1-score) sur l'ensemble de test.
4. Comparez les performances du GMM avec celles des autres modèles (SVM, Voting Classifier).
5. Tracez les frontières de décision du GMM et expliquez pourquoi elles peuvent être différentes des SVM.

Exercice 4 : Optimisation des hyperparamètres

Utiliser des techniques avancées comme *GridSearchCV* pour optimiser les modèles.

1. Quels sont les hyperparamètres principaux d'un SVM que vous pouvez optimiser ?
2. Effectuez une recherche par *GridSearchCV* pour optimiser les paramètres C et kernel du SVM.
3. Optimisez les paramètres du Voting Classifier en ajustant les poids attribués à chaque modèle.
4. Implémentez une optimisation des paramètres du GMM (nombre de composantes, covariance). Évaluez les performances du modèle optimisé.
5. Discutez de l'impact de l'optimisation sur les performances des modèles.

Exercice 5 : Comparaison globale et discussion

Comparer les modèles sur des critères de performance, complexité et adéquation au problème.

1. Comparez les performances globales des modèles en termes de précision, rappel, F1-score et AUC-ROC.
2. Comparez la complexité des modèles (temps d'entraînement, ressources nécessaires).
3. Dans quel contexte un Voting Classifier serait-il préférable à un GMM ou un SVM ?
4. Analysez les avantages et limites des SVM, des GMM et des Voting Classifiers.
5. Proposez une approche hybride en combinant les modèles explorés dans le TP pour des performances optimales.

Remarques :

- URL de l'ensemble de données SMS Spam Collection :

  https: archive.ics.uci.edu/ml/machine-leaming-databases/00228 smsspamcollection.zip

• Utiliser les bibliothèques scikit-leam, matplotlib, et numpy. • Documenter le code de manière claire.

- Explorer les possibilités de personnalisation des modèles (paramètres du noyau, etc.).