Option : Ml OIVM
Module : TP ATDN

TP 3 : Classification binaire avec approches probabilistes

Objectif : Explorer et comparer les approches probabilistes génératives, probabilistes discriminantes, pour la classification binaire en utilisant le jeu de données SMS Spam Collection.

Exercice 1 : Approche Probabiliste Générative : Naïve Bayes

1. Charger et comprendre le jeu de données SMS Spam Collection.
2. Prétraiter les données (nettoyage, sélection des caractéristiques pertinentes).
3. Diviser les données en ensembles d'entraînement et de test.
4. Entraîner un modèle Naïve Bayes (sur l'ensemble d'entrainant).
5. Évaluer les performances du modèle sur l'ensemble de test en utilisant les métriques : la précision, le rappel et le F1-score (prenez le soin de formuler l'écriture mathématique de chaque métrique dans le rapport que vous allez me rendre. Donnez d'autres autres métriques que vous aurez pu utilisez en exploitant la documentation de la librairie).
6. Expliquer l'intérêt d'un modèle génératif par rapport aux autres.

Exercice 2 : Classification avec Complement Naive Bayes

Utiliser le modèle Complement Naive Bayes pour la classification binaire sur des données déséquilibrées et évaluer ses performances.

1- Charger le jeu de données SMS Spam Collection.
2- Nettoyer les données et vectoriser les textes (comme dans les exercices précédents).
3- Former le modèle **Complement Naive Bayes (**Entraîner le modèle sur l'ensemble d'entraînement.)
4- Utiliser des métriques standards (précision, rappel, F1-Score,…).
5- Ajouter une matrice de confusion pour mieux analyser les erreurs.
6- Expliquer votre matrice de confusion.

Exercice 3 : Approche Probabiliste Discriminante : Régression Logistique

Étapes :

1. Charger le même jeu de données.
2. Prétraiter et vectoriser les données de manière similaire à la partie 1.
3. Diviser les données de manière identique.
4. Entraîner un modèle de régression logistique.
5. Évaluer les performances du modèle.
6. Expliquer l'intérêt d'un modèle génératif par rapport aux autres.

Exercice 4 : Comparaison des Performances

1. Créer une table de comparaison des performances (précision, rappel, F1-score).
2. Visualiser les frontières de décision des modèles (optionnel si les données sont transformées en 2D).
3. Comparer les performances des approches en termes de précision, rappel et F1-score.
4. Utiliser des visualisations pour comparer les frontières de décision des modèles (par exemple : un graphique avec des contours de décision).

Discussion :

1. Analyser les résultats obtenus et discuter les avantages et inconvénients de chaque approche.
2. Comparer la complexité des modèles et le temps d'entrainement (avec time ou timeit)..
3. Réfléchir sur la pertinence de chaque approche pour des problèmes de classification spécifiques.
4. Expliquer pourquoi Naïve Bayes est rapide mais peut être limité, tandis que la régression logistique est plus flexible.
5. Quel modèle est le plus adapté si le dataset est déséquilibré ou bruyant ?

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.).