Département IA & SD le 02/12/2025

Section: 4 ING Software

Module : FD1 TP N° 5 et TP N° 6

Ce TP vous demande d'effectuer un projet de **classification de données** de bout en bout en utilisant l'algorithme des **k plus proches voisins (k-NN)**. Le processus se divise en plusieurs étapes.

1) Préparation des données (Preprocessing)

L'objectif est de préparer un jeu de données de référence ("benchmark") pour la classification. Cela implique de :

- **Lire et charger** un benchmark spécifique (comme "Diabète").
- **Nettoyer et prétraiter** les données si nécessaire. Cette étape de **preprocessing** est cruciale pour que l'algorithme fonctionne correctement.

2) Partitionnement du jeu de données

Vous devez diviser le benchmark en deux sous-ensembles :

- **Ensemble d'apprentissage (80%)** : Utilisé pour "apprendre" les relations dans les données.
- **Ensemble de test (20%)** : Utilisé pour évaluer la performance de l'algorithme sur des données qu'il n'a jamais vues.

Cette partition doit être effectuée de manière **aléatoire et sans remise** pour garantir que les ensembles sont représentatifs et distincts.

3) Application de l'algorithme k-NN

Appliquez l'algorithme **k-NN** sur l'ensemble de test, en utilisant l'ensemble d'apprentissage comme référence. Vous devez répéter ce processus en faisant varier la valeur de **k** (le nombre de voisins) de **1 à 10**.

4) Évaluation des performances du classifieur

Pour chaque valeur de k, calculez les mesures de performance suivantes en comparant les prédictions de l'algorithme aux valeurs réelles de l'ensemble de test :

- **Matrice de confusion** : **TP** (Vrais positifs) **TN** (Vrais négatifs) **FP** (Faux positifs) **FN** (Faux négatifs)

- **Mesures dérivées** : **Précision** **Rappel** **F-mesure**

Vous devez afficher ces mesures pour chaque valeur de k (de 1 à 10).

5) Visualisation et analyse des résultats

Dessiner la courbe de précision : Créez un graphique avec les valeurs de k sur l'axe des x et la précision sur l'axe des y.
Déterminer le meilleur k : En vous basant sur la courbe et les mesures, identifiez la valeur de k qui a donné la plus haute précision. Cette valeur est considérée comme le **paramètre optimal** pour le classifieur sur ce jeu de données.

Rapport à remettre le 09/12/2025

6) Application de l'algorithme Naive Bayes

Appliquez l'algorithme **Naive Bayes** sur l'ensemble de test, en utilisant l'ensemble d'apprentissage comme référence.

7) Application de l'algorithme C4.5

Appliquez l'algorithme **C4.5** sur l'ensemble de test, en utilisant l'ensemble d'apprentissage comme référence.

8) Application de l'algorithme SVM

Appliquez l'algorithme **SVM** sur l'ensemble de test, en utilisant l'ensemble d'apprentissage comme référence.

Rapport à remettre le 16/12/2025