Les bases de Machine Learning

2023/2024

II-BDCC 1ere Année

Examen final

Nom & prénom : **Hamza JITTOU**

exercice 1 : Question de cours

1. **Ce que c'est Machine Learning ?**

Réponse :

Le Machine Learning est une branche de l'intelligence artificielle où les ordinateurs apprennent à partir de données pour faire des prédictions ou prendre des décisions, sans être explicitement programmés pour effectuer ces tâches.

2. **Quelle est la différence entre l'apprentissage supervisé et l'apprentissage non supervisé ?**

Réponse :

L'apprentissage supervisé utilise des données étiquetées pour entraîner le modèle à prédire des sorties spécifiques, tandis que l'apprentissage non supervisé utilise des données non étiquetées pour identifier des motifs ou des structures cachées dans les données

3. **Donner une définition aux concepts suivants :**

- Problème de classification :Une tâche d'apprentissage supervisé où le modèle doit assigner des étiquettes ou des catégories à des entrées données.
- Problème de régression : Une tâche d'apprentissage supervisé où le modèle doit prédire des valeurs numériques continues basées sur des entrées données.

4. **Quelles sont les principales métriques à utiliser pour évaluer un modèle de régression ?**

Réponse :

- Moyenne des erreurs quadratiques (MSE)
- Erreur quadratique moyenne (RMSE)
- Erreur absolue moyenne (MAE)
- Coefficient de détermination (R^2)

Certaines des métriques d'évaluation les plus couramment utilisées comprennent la précision, le rappel, le F-score, l'AUC-ROC pour les problèmes de classification, et l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE), le coefficient de détermination (R^2) pour les problèmes de régression.

5. **Quelles sont les principales métriques à utiliser pour évaluer un modèle de classification ?**

Réponse :

- Exactitude (Accuracy)
- Précision (Precision)
- Rappel (Recall)
- F-mesure (F1-score)
- Aire sous la courbe ROC (AUC-ROC)

6. **Quel est le rôle de chaque librairie parmi les librairies suivants :**

- Numpy :Fournit des structures de données et des fonctions pour les calculs numériques efficaces, notamment des tableaux multidimensionnels (ndarray).

- Pandas:Offre des structures de données flexibles et des outils pour la manipulation et l'analyse de données, principalement avec les DataFrames et Series.

- matplotlib:Permet de créer des visualisations graphiques en 2D, comme des graphiques, des histogrammes, et des courbes.
- Sklearn : Fournit des outils pour l'apprentissage automatique, y compris des algorithmes de classification, de régression, de clustering, et des fonctions pour la validation de modèles et la prétraitement des données.

Exercice 2 : Problème de régression

2. Type de problème

a. Écrire l'instruction qui permet de montrer que target est de type numérique.

· **Réponse est déjà donnée dans le point précédent.**

b. Pourquoi il s'agit d'un problème de régression ?

· **Réponse : C'est un problème de régression car la variable cible est une variable numérique continue.**

5. Split le dataset

a. Pourquoi est-il nécessaire de diviser le dataset en training dataset et test dataset ?

· **Réponse : Pour évaluer la performance du modèle sur des données non vues et vérifier sa généralisation.**

b. Quelle fonction utiliser pour diviser le dataset en train (X_train,y_train) et en test (X_test,y_test) ? quels sont ces principaux paramètres ?

· **Réponse : La fonction train_test_split. Les principaux paramètres sont X, y, test_size, et random_state.**

c. Pourquoi il est important de fixer le paramètre random_state ? utiliser random_state=23

· **Réponse : Pour garantir la reproductibilité des résultats en assurant que la division du dataset est la même à chaque exécution.**

6. Model

a. Quelle est la forme mathématique du modèle qui correspond à ce dataset ?

· **Réponse : La forme mathématique est $y = \beta0 + \beta1*x1 + \beta2*x2$.**

3 Partie 2 : création du modèle sans utiliser sklearn

1. L'objectif est de créer notre propre fonction fit(X,y) qui se base sur l'algorithme de gradient descent

a. Ce que c'est l'algorithme de gradient descent ?

· **L'algorithme de gradient descent est une méthode d'optimisation pour trouver les paramètres minimisant une fonction de coût. Il ajuste les paramètres du modèle en suivant le gradient de la fonction de coût pour converger vers une solution optimale.**

b. Quelles sont les principales étapes de l'algorithme ?

· **Initialisation des paramètres : Commencer avec des valeurs initiales pour les paramètres. · Calcul du gradient : Évaluer le gradient de la fonction de coût par rapport aux paramètres. · Mise à jour des paramètres : Ajuster les paramètres en fonction du gradient et d'un taux d'apprentissage. · Répétition : Répéter les étapes 2 et 3 jusqu'à ce que les changements deviennent négligeables ou qu'un nombre maximal d'itérations soit atteint.**

Exercice 3 : Problème de classification

4. Encodage des variables catégorielles

a. Pourquoi la vérification de type de variable est une phase importante ?

· **La vérification est importante pour s'assurer que le modèle peut gérer correctement les types de variables surtout si elles doivent être encodées**