PROJET INTELLIGENCE ARTIFICIELLE

SUJET : METHODE CATBOOST

1 EXPLICATION DU MODEL

1.1 définition :

CATBOOST est une bibliothèque open source hautes performances pour le gradient boosting sur les arbres de décision

1.2 Ces fonctionnalités :

Grande qualité sans réglage des paramètres :

Réduisez le temps consacre au réglage des paramètres, car CatBoost fournit d'excellents résultats avec les paramètres par défaut

Prise en charge des fonctionnalités catégorielles :

Améliorez vos résultats d'entrainement avec CatBoost qui vous permet d'utiliser des facteurs nos numériques, au lieu d'avoir à prétraiter vos données ou à passer du temps et des efforts à les transformer en chiffre

Version GPU rapide et évolutive :

Entrainez votre model sur une mise en œuvre rapide de l'algorithme d'amplification de gradient pour GPU. Utiliser une configuration multi-cartes pour les jeux de données volumineux

Précision améliorée :

Réduise le surajustement lors de la construction de vos model grâce a un nouveau schéma d'amplification du gradient

Prédiction rapide :

Appliquez votre modèle forme rapidement et efficacement, même aux taches critiques en termes de latence, à l'aide de l'applicateur de modèles de CatBoost

A propos :

CatBoost est un algorithme de gradient boosting sur les arbres de décision .il est développé par des chercheurs et ingénieurs de Yandex et est utilise pour la recherche, les systèmes de recommandation, l'assistant personnel, les voitures autonomes les prévisions météorologiques et de nombreuses autres taches chez Yandex et dans d'autres entreprises, notamment le CERN, Cloudflare, Careem taxi.il est en open-source et peut être utiliser par n'importe qui.

2 DIFFERENCE ENTRE CATBOOST ET LES AUTRES METHODES DE ML

CatBoost	KNN	Arbre de Décision	Naive Bayes	Régression logique
CatBoost est un algorithme d'apprentissage automatique à source ouverte récemment de Yandex. Il peut facilement s'intégrer à des frameworks d'apprentissage en profondeur tels que TensorFlow de Google et Core ML d'Apple. La meilleure partie de CatBoost est qu'il ne nécessite pas de formation approfondie sur les données comme les autres modèles de ML, et peut fonctionner sur une variété de formats de données; pas nuire à sa robustesse. Assurez-vous de bien gérer les données manquantes avant de procéder à l'implémentation. Catboost peut traiter automatiquement les variables	Il peut être utilisé à la fois pour les problèmes de classification et de régression. Cependant, il est plus largement utilisé dans les problèmes de classification dans l'industrie. K plus proches voisins est un algorithme simple qui stocke tous les cas disponibles et classe les nouveaux cas par un vote majoritaire de ses k voisins. Le cas assigné à la classe est le plus courant parmi ses K voisins les plus proches mesurés par une fonction de distance. Ces fonctions de distance peuvent être la distance euclidienne, Manhattan, Minkowski et Hamming. Les trois premières fonctions sont utilisées pour la fonction	L'arbre de décision est un algorithme qui se base sur un modèle de graphe (les arbres) pour définir la décision finale. Chaque nœud comporte une condition, et les branchements sont en fonction de cette condition (Vrai ou Faux). Plus on descend dans l'arbre, plus on cumule les conditions. L'image ci-dessus illustre ce fonctionnement	Naïve Bayes est un classifieur assez intuitif à comprendre. Il se base sur le théorème de Bayes des probabilités conditionnelles. Naïve Bayes assume une hypothèse forte (naïve). En effet, il suppose que les variables sont indépendantes entre elles. Cela permet de simplifier le calcul des probabilités. Généralement, le Naïve Bayes est utilisé pour les classifications de texte (en se basant sur le nombre d'occurrences de mots).	La régression logistique est une méthode statistique pour effectuer des classifications binaires. Elle prend en entrée des variables prédictives qualitatives et/ou ordinales et mesure la probabilité de la valeur de sortie en utilisant la fonction sigmoïd (représentée dans la photo). On peut effectuer la classification multi-classes (par exemple classifier une photo en trois possibilités comme moto, voiture, tramway). En utilisant la régression logistique et la méthode un-contre-tous (One-Versus-All classification). La régression logistique permettra de répondre à des

catégorielles sans afficher l'erreur de conversion de type, ce qui vous aide à mieux vous concentrer sur le réglage de votre modèle plutôt que sur le tri des erreurs triviales.	continue et la quatrième (Hamming) pour les variables catégorielles. Si K = 1, alors le cas est simplement assigné à la classe de son voisin le plus proche. Parfois, choisir K s'avère être un défi lors de l'exécution de la modélisation kNN			problèmes comme : Est-ce que le client est solvable pour lui accorder un crédit ? Est-ce que la tumeur diagnostiquée est bénigne ou maline ?