

XGBoost – Qu’est-ce que c’est ?

XGBoost est un algorithme d'apprentissage machine basé sur un arbre de décision d'ensemble et utilise un cadre de renforcement de gradient. Les réseaux neuronaux artificiels sont généralement plus performants que d'autres cadres ou algorithmes lorsqu'il s'agit de prédire des problèmes avec le texte, les images et d'autres données non structurées. Les choses sont très différentes lorsqu'il s'agit de données tabulaires/structurées de taille moyenne ou petite. Pourquoi ? Parce que les algorithmes basés sur des arbres de décision sont sans doute la meilleure solution.

L'algorithme XGBoost est le résultat d'un projet de recherche organisé à l'Université de Washington. Carlos Guestrin et Tianqi Chen ont présenté leur article lors de la conférence SIGKDD de 2016. Depuis lors, XGBoost est devenu un algorithme réputé qui a révolutionné le paysage de l'apprentissage machine.

Depuis son introduction, cet algorithme a reçu des tonnes de crédits, dont plusieurs concours Kaggle. Il a également été reconnu pour son influence sur diverses applications industrielles de pointe. Par conséquent, un groupe important de scientifiques spécialisés dans les données contribue à différents projets open source liés à XGBoost. GitHub compte plus de trois cents contributeurs, et trente-six cents engagements à ce jour. Découvrez comment cet algorithme se distingue des autres.

Comment XGBoost se démarque

Un vaste éventail de demandes

XGBoost vous permet d'utiliser un large éventail d'applications pour résoudre des problèmes de prédiction, de classement, de classification et de régression définis par l'utilisateur.

Facilement portable

Il fonctionne sans problème sous OSX, Linux et Windows.

Variété de langues

XGBoost supporte la plupart des langages de programmation, notamment Julia, Scala, Java, R, Python, C++.

Intégration avec le Cloud

L'algorithme d'apprentissage machine XGBoost prend en charge les grappes de fils, l'azur et l'AWS. Il fonctionne aussi incroyablement bien avec d'autres écosystèmes Spark, Flink et divers.

Construire une intuition XG Boost

L'une des raisons pour lesquelles les arbres de décision sont si populaires est que leur visualisation est assez simple dans leur forme la plus élémentaire. Leurs algorithmes sont également faciles à interpréter, mais créer une intuition pour la prochaine génération d'algorithmes (basés sur les arbres) peut être un peu compliqué. Voici une analogie pour comprendre les algorithmes basés sur les arbres et leur évolution.

Imaginez que vous êtes un responsable du recrutement, en train de parler à plusieurs candidats hautement qualifiés. Chaque étape de l'évolution des algorithmes arborescents est comme une partie du processus d'entretien.

Arbre de décision

Chaque responsable du recrutement suit un ensemble de critères, notamment l'expérience, la formation et les résultats des entretiens. Il serait juste de comparer un arbre de décision à un responsable du recrutement qui interroge les candidats en fonction de ses critères

Mise en sac

Plusieurs personnes sont désormais en charge des interviews, et tous les membres du jury ont une voix. L'agrégation ou la mise en sac bootstrap consiste à regrouper les contributions de chaque intervieweur pour prendre une décision définitive à l'aide d'un processus de vote impartial.

Forêt aléatoire

Le Random Forest est un algorithme de mise en sac qui sélectionne de manière aléatoire un sous-élément de caractéristiques. En d'autres termes, chaque intervieweur testera le candidat sur la base de qualifications choisies au hasard (par exemple, des entretiens techniques pour tester les compétences en programmation)

Quel est le secret des excellentes performances de XGBoost ?

Les machines de gradient boosting et XGBoost sont des méthodes d'arbre d'ensemble. Elles utilisent l'architecture à descente de gradient pour appliquer le principe de boost des apprenants pauvres (CART). Extreme Gradient Boosts améliore considérablement le cadre GBM en mettant en œuvre des améliorations algorithmiques et une optimisation des systèmes.

XGBoost pourrait être l'avenir de l'apprentissage machine

Les scientifiques doivent tester chaque algorithme pour identifier la meilleure option. En outre, choisir le bon algorithme, bien qu'excellent, ne suffit pas. Le choix de la bonne configuration pour un ensemble de données particulier est également crucial. De plus, il faut

tenir compte de plusieurs éléments pour choisir le meilleur algorithme. Faites attention à des éléments comme la mise en œuvre, l'explicabilité et la complexité de calcul pour déterminer l'algorithme qui vous conviendrait le mieux.

Enfin, l'apprentissage machine est une vaste alimentation, et vous trouverez des tonnes d'alternatives à l'Extreme Gradient Boost. Cat Boost et Light GBM sont des alternatives récentes qui donnent des résultats impressionnants. Cela dit, XGBoost règne en maître en termes d'explicabilité, de pragmatisme, de flexibilité, de performances et de bien d'autres choses encore. Attendez-vous à ce que cet algorithme d'apprentissage machine reste au top jusqu'à ce qu'un digne concurrent arrive et bouscule tout.