Clément

Arki

BTS SIO 2

Concepts importants du machine Learning

1. Concept fondamental

Le Machine Learning est la science du développement d'algorithmes et modèles statistiques que les systèmes informatiques utilisent pour effectuer des tâches sans instructions explicite. C'est une manière de modéliser des phénomènes, dans le but de prendre des décisions stratégiques. Le Machine Learning existe une relation mathématique entre toute combinaison de données d'entrée et de sortie.

2. Données

Les données est le nerf de la data science. Dans les données il y a :

- Les bases de données
- Les données brutes tel que
    - Le texte
    - Les images (et vidéos)
    - IoT (Internet des objets)

3. Les 4 types d'apprentissage

Les 4 types d'apprentissage sont :

- Les données
- La tâche spécifique à accomplir
- L'algorithme d'apprentissage
- L'analyse d'erreur (ou mesure des performances du modèles)

4. Prétraitement des données

Le prétraitement des données est l'étape pendant laquelle les données brutes sont nettoyées et structurées en vue de l'étape suivante du traitement des données. Pendant cette phase de préparation, les données brutes sont vérifiées avec soin afin de déceler d'éventuelles erreurs.

5. **Choix du modèle**

Un modèle Machine Learning est un fichier qui a été entraîné pour reconnaitre certains types de modèles

Comme le choix du modèle par exemple est un modèle sous-jacent (ou modèle statistique) car il représente une contrainte de « forme » non dépendante des données.

Par exemple, si je décide d'effectuer une régression linéaire, je contrains mon modèle à avoir la forme d'une droite et pas d'un cercle. Malgré tous les efforts, on ne pourra pas obtenir d'autre forme qu'une droite pour modéliser les données, c'est donc un choix très important.

6. **Entraînement et validation**

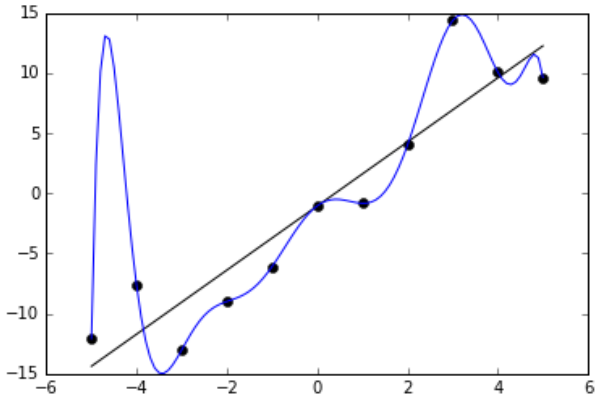
L'entrainement est qu'un modèle est exposé à un ensemble de donnée d'entraînement pour apprendre à faire des prédictions. Tandis que la validation évalue la capacité à faire des prédictions précises sur de nouvelles données.

7. **Evaluation du modèle**

L'évaluation du modèle en Machine Learning consiste à mesurer la performance, la précision et l'efficacité d'un modèle sur des données de test ou de validation afin de déterminer sa capacité à prendre des décisions précises et à généraliser correctement sur de nouvelles données. Pour l'évaluation du modèle, il y a la validation croisée.

8. **Surapprentissage (overfitting)**

Le surapprentissage (overfitting) est le fait que le modèle qu'on a choisi est trop collé aux données d'entraînement. C'est un problème classique de data science, lorsqu'on choisit un modèle trop "flexible", c'est-à-dire avec une complexité trop élevée qui prend aussi en compte le bruit du phénomène. C'est aussi un comportement indésirable d'apprentissage automatique qui se produit lorsque le modèle d'apprentissage automatique fournit des prédications précises pour les données d'entrainement mais pas pour les nouvelles données.



9. Hyperparamètres

Un hyperparamètre est un paramètre qui est utilisé pour configurer un modèle Machine Learning. Contrairement aux paramètres du modèles, qui sont appris à partir des données d'entraînement, les hyperparamètres doivent être définis par l'utilisateur avant de commencer l'entraînement du modèle.

10. Réglage des hyperparamètres

Le réglage des hyperparamètres permet aux spécialistes des données d'ajuster les performances du modèle pour obtenir des résultats optimaux. Ce processus est une partie essentielle du Machine Learning, et le choix des valeurs appropriées des hyperparamètres est crucial pour le succès.

11. Validation croisée

C'est la validation croisée qui permet donc d'évaluer un modèle de machine learning en ayant la moyenne des performances et l'erreur type sur chacun des folds (c'est-à-dire on coupe donc les données connues en parties égales dans la mesure du possible) ou évaluant les prédications faites sur l'ensemble des données. Il y a aussi l'ensemble de données de test, les métriques d'évaluation, l'analyse des résultats, la correction et l'ajustement et l'interprétation des résultats.

12. Généralisation

La généralisation désigne la capacité du modèle à pouvoir effectuer des prédictions robustes sur des nouvelles données. Lorsqu'on veut généraliser un modèle, on a besoin qu'il overfit pas et qu'il n'underfit pas, qu'il soit pile entre les deux.

13. Déploiement

Le processus de déploiement de modèles de Machine Learning requiert plusieurs entrées. Il a un accès à l'ensemble des données brutes nécessaires à l'entraînement du modèle. Il a un script d'entraînement développé par des Data Scientists. Et il a une plateforme pour réaliser l'entraînement de modèles (Cloud)

14. Ethique

L'éthique pour la Machine Learning dans l'IA consiste pour l'essentiel à remettre en question, à enquêter en permanence et à ne jamais considérer comme acquises les technologies qui s'imposent rapidement à la vie humaine.

15. Apprentissage continu

L'apprentissage continu cherche à apporter des solutions à ce problème en conférant à une méthode de Machine Learning la capacité de s'adapter.
Quand une Machine Learning doit être intégrée au sien d'un métier, la faculté à se maintenir à jour en poursuivant son apprentissage est très souvent requise. Mais l'apprentissage continu se heurte à deux difficultés, l'oubli catastrophique et enfin le lien entre l'architecture d'une méthode de Machine Learning et le nombre de données qu'elle peut assimiler

16. Domaines d'application

Le domaine d'application pour le Machine Learning concerne tous les secteurs d'activité, notamment l'industrie, le commerce, la santé et les sciences de la vie, le tourisme et l'hôtellerie, les services financiers, l'énergie, les matières premières et les services publics.