Sommaire

A propos

Ce document présente les différentes étapes par lesquelles vous devriez impérativement passer quel que soit votre projet. Voyez cette notice comme un fil conducteur si vous ne savez pas par où commencer et/ou si vous êtes perdus en cours de route. A la fin de cette note vous trouverez aussi quelques informations sur ce qui est attendu de votre part et dont votre note sera directement dérivée. Nous fournissons également quelques conseils pour bien commencer et conduire votre projet ainsi que pour bien monter en compétence avec ce module.

Première étape : identifier et définir le problème

La première étape va consister à passer de l'intitulé du projet que vous avez choisi à une définition plus rigoureuse.

Commencer par demander quelles sont les entrées et les sorties du modèle. Par exemple, si le sujet est la classification d'images de chiens et de chats alors l'entrée est une image qui contient soit un chat, soit un chien et la sortie sera "chien" ou "chat" selon ce que contient l'image. En général, on utilise plus souvent 0 ou 1 plutôt qu'un mot pour une question pratique.

Dans le cas d'un système de recommandations, l'entrée pourrait être un historique de transactions et un utilisateur et la sortie un produit ou une liste de produits susceptibles d'intéresser l'utilisateur en question.

Il est également important de savoir identifier si le problème est une classification ou une régression car cela impactera fondamentalement l'évaluation du modèle.

Deuxième étape : comprendre les données

Maintenant que l'objectif est clairement établi, passons à la compréhension et l'exploration des données. Cette étape est extrêmement importante, assurez-vous avant d'aller plus loin de savoir correctement répondre aux questions suivantes :

- Quantité de données :
    - Quel est le format des données ?
    - Quelle est la taille de la base de données (en nombre de lignes et de colonnes, en nombre d'images, etc.) ?
- Qualité des données
    - Les données comprennent-elles des caractéristiques pertinentes pour ma problématique ?
    - Quels sont les types de données présents (symbolique, numérique, etc.) ?
    - Avez-vous calculé des statistiques de la base pour les attributs-clés ? En quoi cela va-t-il permettre d'éclaircir la problématique ?
    - Est-il possible d'identifier les attributs pertinents ?

Voici d'autres questions que vous pouvez vous poser (suivant votre problématique, certaines de ces questions peuvent ne pas avoir du sens) :

- Quels sont les attributs qui semblent sans intérêt et peuvent être exclus ?
- Le nombre de données permet-il de tirer des conclusions pouvant être généralisées ou d'effectuer des prévisions précises ?
- Les attributs sont-ils trop nombreux pour la méthode de modélisation choisie ?
- Avez-vous envisagé le mode de traitement des valeurs manquantes dans chacune de vos sources de données (voir tableau en fin de cette partie) ?
- Avez-vous identifié des attributs manquants et des champs vides ? Si oui, ces valeurs manquantes ont elles une signification ?
- L'orthographe présente-t-elle des incohérences pouvant engendrer des problèmes lors de fusions ou de transformations ultérieures ?
- Avez-vous exploré les écarts afin de déterminer s'il existe des valeurs aberrantes ou des phénomènes à analyser plus en profondeur ?
- Avez-vous vérifié la plausibilité des valeurs ? Relevez les conflits apparents (par exemple des images ne contenant rien, des tweets vides etc.).
- Avez-vous envisagé d'exclure les données sans impact selon vos hypothèses ?
- Avez-vous utilisé des graphiques exploratoires pour clarifier les attributs-clés ?
- Savez-vous quels sont les attributs à filtrer ou à sélectionner, et les sources de données à fusionner ?

Problème posé par les données	Solution possible
Données manquantes	Excluez les lignes ou les caractéristiques, ou insérez une valeur estimée dans les blancs.
Erreurs dans les données	Procédez de manière logique pour découvrir manuellement les erreurs et les corriger, ouexcluez les caractéristiques.
Codage des incohérences	Décidez d'une méthode de codage unique, puis convertissez et remplacez les valeurs.
Métadonnées erronées ou manquantes	Examinez manuellement les champs suspects et recherchez la signification correcte.

Comment allez-vous évaluer votre modèle ?

Une fois que vous savez ce que devra faire votre modèle (classer, faire une régression, etc.). Vous devriez déjà être en mesure de choisir comment évaluer votre modèle (indépendamment du choix de ce dernier). Cela passe par le choix d'une métrique adaptée. Souvent pour un problème de régression on utilisera la MSE (Mean Square Error) et la precision pour les problèmes de classification.

Attention veillez que la métrique choisie correspond bien à votre problème. Vous trouverez ici les métriques les plus utilisées ainsi que le type de problème associé (classification, régression, etc.).

La modélisation et le choix d'un modèle

Une analogie

Entrainer un modèle, c'est comme accorder une guitare, vous faites vibrer les cordes (les vibrations, ce sont vos données d'entrée.) en attendant un certain son en retour (ce sont les données de sortie.). Au début la guitare n'est pas accordée alors les sons produits ne sont pas ceux attendus, vous comparez à l'oreille ce que vous entendez à ce que vous souhaitez entendre et en fonction de cette différence, vous allez régler les boutons (les paramètres de votre modèle).

Par analogie, pour un modèle d'apprentissage, vous allez fournir des données en entrée ainsi que la sortie attendue (souvent dénotés "X" et "y" respectivement). Au début, le modèle n'est pas adapté (en général, ses paramètres sont initialisés aléatoirement.), à l'aide d'un algorithme, le modèle va accorder ses paramètres, c'est-à-dire tenter de faire correspondre sa sortie avec la sortie attendue en fonction de l'entrée que vous lui fournirez en jouant sur la valeur de ses paramètres internes.

Il faut bien distinguer le modèle (la guitare) de l'algorithme qui va ajuster les paramètres (le guitariste), souvent cette distinction est floue, car les bibliothèques qui fournissent les modèles utilisent des algorithmes par défaut.

En général, quand on parle d'IA, on regroupe le modèle et son algorithme d'apprentissage.

Dans l'exemple de la régression qui vous a été présentée dans le premier cours, le modèle est une droite (hyperplan), l'algorithme qui permet d'ajuster les paramètres de cette droite est la méthode des moindres carrés. On peut utiliser ce même modèle (la droite) avec d'autres algorithmes, par exemple, on peut générer aléatoirement les paramètres de la droite (c'est très inefficace, mais c'est un bon exemple pour comprendre le principe).

Comment sait-on si l'entraînement est bon ?

Lorsque l'on entraîne un modèle, il faut suivre quelques règles. L'une des plus importante consiste a séparer votre jeu de données en deux ensembles, un pour l'entraînement et un autre pour le test (d'autres partitionnements existent). Le modèle sera accordé grâce aux données d'entraînement et ses performances seront évaluées sur l'ensemble de test. Pourquoi ne pas entraîner et évaluer directement sur toutes les données ? Pour la même raison que vos partiels ne contiennent pas les mêmes exercices que dans les TD. En effet, on souhaite éviter l'apprentissage "par cœur".

Le modèle doit être capable de généraliser son apprentissage à des données qu'il n'a pas vues durant son entraînement. Imaginez une voiture autonome qui ne peut conduire que sur les routes qu'elle connaît ou un algorithme qui ne pourrait conseiller uniquement que des films que vous avez déjà vus.

Pour éviter ce cas de figure, on entraîne et on évalue le modèle sur deux ensembles distincts. Il n'y a pas de règle absolue sur la manière dont les données doivent être découpées, en général, on prend autour de 80% des donnes pour l'entraînement et le reste pour le test.

Quelques questions à se poser

Lorsque vous décideriez quel modèle utiliser, vérifiez les points suivants pour savoir s’ils ont une incidence sur votre choix :

- Le modèle exige-t-il que les données soient divisées en ensembles de test et d’apprentissage ?
- Avez-vous suffisamment de données pour produire des résultats fiables avec un modèle donné ?
- Le modèle exige-t-il un certain niveau de qualité des données ? Vos données actuelles répondent-elles à ce niveau ?
- Le type de vos données est-il approprié pour un modèle spécifique ?
  Si ce n’est pas le cas, pouvez-vous effectuer les conversions nécessaires ?
- Quelles données seront utilisées pour tester les modèles ?
- Avez-vous partitionné les données en ensembles d’apprentissage/de test ?

Au final qu'est-ce qui est attendu ?

1. À partir de l'intitulé, identifier le type de problème, l'entrée et la sortie de modèle à produire
2. Ouvrir, afficher, comprendre ses données.
3. Choisir une méthode pour évaluer son modèle.
4. Choisir un modèle adapté à son problème et comprendre dans les grandes lignes (être

capable en 2-3 phrases de faire comprendre à vos collègues) le fonctionnement du modèle et le mécanisme d'apprentissage sous-jacent.
5. Appliquer le modèle à ses données et commenter les résultats.
6. BONUS si vous comparez plusieurs modèles en sachant identifier les avantages et les inconvénients de chacun.

Quelques conseils pour une bonne maîtrise

- Ne brûlez pas les étapes : assurez-vous d'avoir compris toutes les parties de tous les TPs avant de commencer, prenez le temps de comprendre vos données avant de chercher à appliquer des modèles.
- Assurez vous de comprendre la différence entre régression et classification et d'avoir compris dans quelle catégorie se situe votre problème.
- Assurez-vous d'avoir compris le mécanisme d'entraînement d'un modèle notamment pourquoi on sépare les données en deux ensembles (train/test).
- Commencez simple, par exemple si vos données possèdent plusieurs attributs, essayer avec 1-2 au début et avec une régression, ensuite, ajoutez des attributs, utilisez des modèles plus complexes, etc. Mieux vaut une solution simple qui fonctionne qu'une méthode complexe que vous ne maîtrisez pas.
- Faites des recherches, il se peut que votre problématique ne soit ni une classification ni une régression classique (exemple recommandation de films, prédiction du cours de la bourse, etc.) mais des ressources en ligne sont extrêmement nombreuses, apprenez à chercher et à trier l'information.
- Venez aux séances avec des questions.