Api la Data Science pour tous -- Jour 1

Introduction à la Data Science

[TOC]

Présentation

Support : https://www.hds.utc.fr/~tdenoeux/dokuwiki/_media/en/utc_sept_2018.pdf * Sébastien Destercke * Enseignant chercheur au GI - FDD - CNRS

L'Intelligence artificielle (IA)

Reproduire des raisonnements humains de manière automatique. C'est un domaine très vaste qui regroupe plus ou moins de nombreuses disciplines (vision par ordinateur, théorie des jeux etc.).

Fondements théoriques de l'informatique par Alan Turing, inventeur du test de Turing : "Suis-je en train de parler à une machine ou non ?".

Claude Shannon, acteur très important aussi.

- IA symbolique : ensemble de règles définies sous forme logique avec des symboles etc.
- IA forte : développer une machine consciente capable de comportements intelligents. Ce type d'IA reste très peu probable à l'heure actuelle.
- IA faible : Ce qu'on fait de nos jours. Automatiser le plus possible de raisonnements complexes comme le traitement de la parole (Siri par Apple et Amazon Alexa par exemple) ou encore la reconnaissance d'obstacles.

Dans les années 50, on faisait de l'IA symbolique (avec des règles logiques pas forcément très complexes). Ensuite il y a eu l'émergence des systèmes experts. C'était un peu de l'IA mais on n'appellait pas ça comme ça.

- Début des années 90 : Réseaux de neurones (Yann Le Cun)

Maintenant : C'est à la mode, les gens croient que l'IA va tout résoudre.

Apprentissage automatique (Machine Learning)

C'est une branche de l'IA. La machine apprend à partir d'observations.

L'un des challenges les plus connus est le fait de reconnaître des chiffres écrits à la main.

A l'intersection entre statistiques, informatique et science des données (prétraitements).

Applications : recherche Google, recommandations, voitures autonomes (--> Heudiasyc) etc.

Apprentissage supervisé vs apprentissage non supervisé

Régression logistique

Observations -> modèle Tâche à résoudre ? Comment modéliser ?

Ex : infarctus du myocarde - Quelle proba de l'avoir -> analyser âge & taux de chloestérol - Approche expert -> règles Si .. Alors - Si âge > 60 et ldl > 10 Alors risque élevé - Si 50 < âge et 8 < ldl <= 10 Alors risque moyen - Difficile à faire à la main - Mais système interprétable facilement par médecin - Régression - Trouver la droite/modèle qui minimise la somme des erreurs - Choisir les paramètres de la droite (w0, w1, w2) - Méthode de l'entropie croisée - Optimiser une fonction convexe (ici la fonction d'erreur) -> descente de gradient - Non-convexe -> gradient stochastique/recuit simulé

Probabilité d'erreur -> Il ne faut pas simplement minimiser le taux d'erreur

Ex : reconnaissance d'expressions

- Différente expressions simulées (joie, peur, dégoût etc)

Linéaire au non-linéaire

Images d'animaux -> couleurs, très variable, linéaire devient difficile

Trouver l'espace de représentation (cartésien/polaire/...)

Réseaux de neurones profonds

Réseau de neurones (perceptron) -> somme de paramètres (xi)*poids (très simplifié : 2 classes possibles)

Jeu de poids -> somme -> fonction d'activation -> 0 ou 1

- limité donc amélioration avec fonction dérivable

Rétropropager la dérivé de l'erreur

Aujourd'hui -> réseaux multicouche (multiplicité des paramètres)

Il faut beaucoup de données pour que ces modèles soient efficaces

- Données images -> transformation de données (ex : décaler le sujet dans l'image)

Choix du modèle de représentation (droite, 2nd degré etc.) en fct :

- Des données d'entrée (qte, répartition..)
- De l'exploitation en sortie (complexité/exactitude souhaité)
- Peu de paramètres : beaucoup d'erreurs
- Trop de paramètres : impossible de dégager les tendances/patterns dans les données inconnues
- Trouver le juste milieu
- Fonction idéale inaccessible, on veut donc s'en rapprocher le plus possible

Erreur : test - apprentissage

- Point critique d'erreur "optimum" puis divergence entre test et apprentissage -> "surapprentissage"

Réseaux de convolution

Technique maîtrisée aujourd'hui

Marche bien pour la classification d'images

Convolution : Suite de transformations de l'espace pour se ramener à un réseau linéaire

Chaque couche du réseau se spécialise (coins, lignes, détails)

Conclusion

Parmi les grandes questions actuelles : causalité/raisonnement, les techniques symboliques, quantification des incertitudes (exemple de l'erreur "panda"/"cheval")

Modèles reproduisent les régularités statistiques Si les données ont un biais éthique -> sera reproduit par le modèle...

à l'UTC : - GI FDD/ICSI - UVs : SY02 ; SY09; SY27; RO04; IA0[1/2] - Master & recherche