Cours de Data Science

Marc Tommasi

6 septembre 2022

Outline

Outline

Préambule

- Évaluation en CCI
  - Une note finale
  - Moyenne d'une note de Projet et d'un contrôle avant la Toussaint
  - Moyenne de ces deux notes
  - Seconde chance à partir des TP/TD rendus.
- Cours moodle : https://moodle.univ-lille.fr/course/view.php?id=17020
  - Groupe 1 : wx9j5q
  - Groupe 2 : d3dsuw

Objectifs

- Présentation de quelques méthodes, algorithmes fréquemment rencontrés en science des données
- Introduction à l'apprentissage machine
- Introduction de quelques notions théoriques du domaine
- Aspects pratiques
- Vous ne serez pas des experts à la fin du cours !

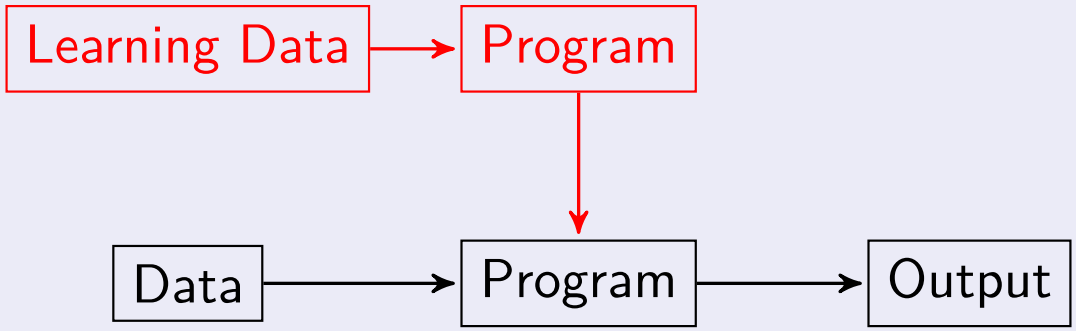
BI vs ML

- BI (business intelligence, informatique décisionnelle)
  - On gère des données massive
  - On étudie la disponibilité, l'accès aux données, la modélisation des données (flocon, étoile), leur stockage (datawarehouse, datamart,...)
  - On conçoit des tableaux de bords (statistiques élémentaires, agrégats,...)
  - On navigue dans ces données (drill up, down,...)
  - On réalise parfois des analyses de tendance
- ML
  - On conçoit des modèles prédictifs, on génère des programmes, à partir de données

Programmation, IA et ML

Programmation

IA et Machine Learning



- IA Classique systèmes experts, systèmes à base de règles, représentation des connaissances, raisonnement logique,. . .
- Machine Learning : Approches complètement dirigées par les données

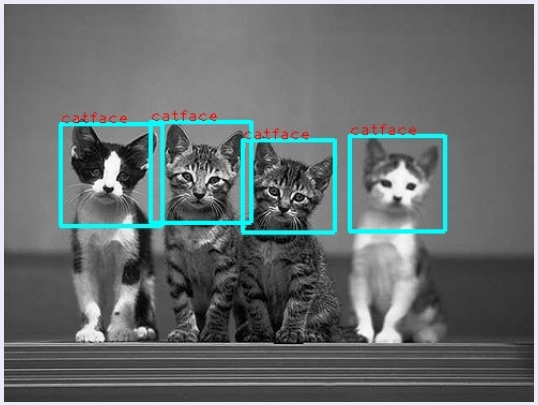
Machine Learning

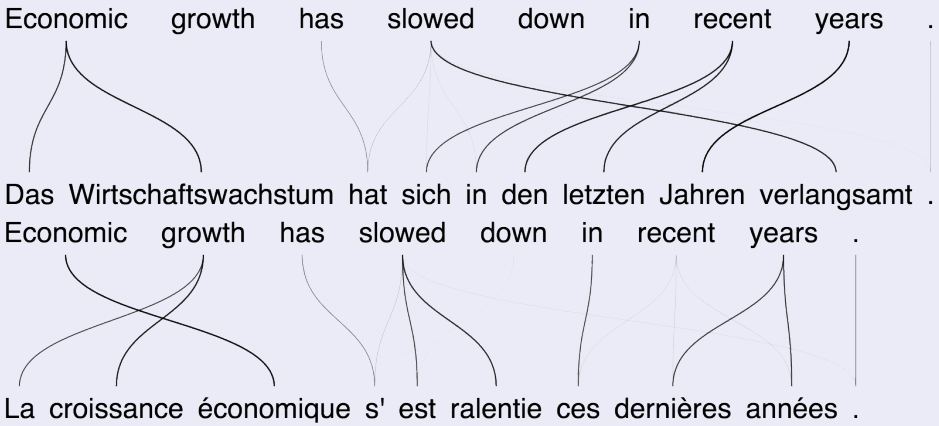
- ML provides a computer with the ability to do certain tasks without being explicitly programmed for it (Arthur Samuel, 1959)
- This is done by learning from data
- Multidisciplinary field : computer science, statistics, optimization
- ML is fueling the current progress in AI
- Processus d'apprentissage est automatisable

Secteurs d'activité

- Services : banques, assurances, commerce et distribution,
  - des tableaux de bord à la prise décision
  - généralisation à de nouveaux services
- Agriculture : encore peu développé, nombreux efforts initiatives actuellement
- Industrie : incitations avec l'industrie 4.0

Exemples d'applications





- Requires large amounts of data (potentially *sensitive, personal data*)

Types de données et Applications

- Images : astronomie, agriculture, météo, archéologie, authentification (reconnaissance faciale), médecine (IRM, ...), OCR, voitures autonomes,...
- Texte : NLU, Génération, Spam, textes médicaux, traduction,
- Son : MIR, STT, Chatbots, sous-titrage vidéo, traduction, reconnaissance et authentification,...
- Données génétiques
- Données de capteurs : transport, robotique, maintenance prédictive, météo,...
- Jeux : Go, échecs, jeux vidéo, ...
- Données du web : recommandation, etc...

Quand passer à de l'apprentissage automatique ?

- Quand la tâche est routinière mais pas explicitement définissable facilement. Trop difficile à programmer directement comme conduire une voiture, reconnaître une image, ou traduire une parole en texte.
- Quand les données sont trop massives et la combinatoire trop importante : prédire la météo, analyser des données génomiques, ... retrouver des motifs dans les données massives en règle générale.
- Quand on doit s'adapter. Les programmes écrits « à la main » sont trop rigides. Pensez à tous les programmes qui demandent une personnalisation (reco de la parole,..) ou doivent évoluer vite (spam).

Histoire, échecs et succès de l'IA

- Dès l'apparition de l'ordinateur !
- Articles importants dans les années 50 et 60
- Période un peu froide à partir de la fin des années 60
- Systèmes experts dans les années 70-80
- Essor du machine learning, apprentissage statistique à partir des années 90
- Succès dans la vision par ordinateur, les jeux, ... par les techniques d'apprentissage profond

Difficultés de l'IA et du ML en particulier

- attaques
- robustesse
- données personnelles
- équité
- énergie
- confiance
- interprétation
- cadre légal
- acceptation sociale et peurs
- différence ML et automatisation (programmation)

Conclusion

- Progrès assez spectaculaires récents
- Passage récent dans l'industrie et le grand public,
- Technologies pas toujours mûres, en constante évolution,
- Besoin de personnes qualifiées, qui s'adapteront aux évolutions certaines
- De nombreuses questions de société.

Outline

Introduction

- Machine Learning : convertir expérience en connaissance et expertise.
- Citation de Mitchell :

  A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on task T, measured by P, improves with experience E.

- entrée : l'expérience est représenté par des données d'apprentissage
- sortie : l'expertise est un autre programme qui réalise une tâche
- Questions :
  - Quelles données en entrée ?
  - Comment automatiser l'apprentissage ?
  - Quand dire qu'on apprend, comment évaluer le succès ?

Apprentissage par cœur et généralisation

- Détection du spam : on peut mémoriser tous les spam qui ont été étiquetés comme tels. A-t-on appris ?
- besoin de généralisation, par exemple repérer des mots qui sont témoins d'un spam
- Importance d'information a priori : on n'apprend pas sans un certain biais (No free lunch theorem)

Type d'apprentissage

- Expérience : données d'entraînement pour l'apprentissage ; Mesure de la performance sur le test.
- Supervisé, non supervisé, par renforcement :
  - En non supervisé train et test n'apportent pas plus d'information,
  - supervisé et par renforcement, le train a plus d'information que le test
- Actif/passif : l'apprenant influence ou pas l'environnement avec lequel il interagit.
- Avec teatcher ou pas : comme l'école ou comme un scientifique face à la nature. NB : le teatcher peut être un adversaire
- Online ou batch : quand doit on prendre la décision ?
- Basé sur un modèle ou sur les instances des données (paramétrique, non paramétrique)

Défis

- Obtenir des données
- Obtenir de bonnes données
  - biais de sampling
  - outliers
  - données non représentatives
- Overfitting
- Underfitting
- Évaluation de l'erreur
- Sélection de modèle ou d'algorithme, mais « no free lunch » !