LU3I026 - Science des données

Introduction

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

Sorbonne Université

2025-2026

Organisation générale

Cours: début lundi 19 janvier

- Lundi 8h45–10h30: amphi ?

TD/TME: à partir du 26 janvier

Groupe 1:

- lundi 10h45–12h30: salle ?
- lundi 14h–15h45: salle ?

Groupe 2:

- mercredi 14h–17h45: salle ?

Groupe 3:

- vendredi 8h45h–112h30: salle ?

Moodle

- Vous serez bientôt inscrits

Intervenants

Cours

- Olivier Schwander
  <olivier.schwander@sorbonne-universite.fr>
- Christophe Marsala <Christophe.Marsala@lip6.fr>

TD/TME

- Christophe Marsala <Christophe.Marsala@lip6.fr>
- Jean-Noël Vittaut <jean-noel.vittaut@lip6.fr>
- Maxellende Julienne <julienne@isir.upmc.fr>

Évaluation

- Epreuve 1 - E1 (cE1 = 25%) : **contrôle écrit de mi-semestre** (mars)
- Epreuve terminale - ET (cET = 40%): **examen terminal** (mai)
- Epreuve 2 (persistante) - E2 (cE2 = 35%) : **projet final** (rendu + soutenances)
- Note finale Session 1 : S1 = cE1 * MAX(E1, ET) + cE2 * E2 + cET * ET
- Note finale Session 2 : S2 = MAX(S1, cE2 * E2 + (cE1 + cET) * SC)

Projet final

- Persistant: compte même en seconde session
- Pas de max
- **IMPORTANT**

Séances de cours

Pendant le cours

- Y assister, **bien évidemment**
- Prendre des notes
- Slides en ligne au début du cours
- Pas de rappels en TD/TME
- **Attention, des morceaux au tableau**

Avant les TD/TME

- Mettre les notes au propre, et les relire
- Revoir les exemples et les exercices
- Avoir disponible ses notes des séances précédentes

Séances de TD/TME

Pas de changement de groupe possible sauf bonne raison

TD

- Pas forcément toutes les semaines
- Travail sur feuille
- Pas d'ordinateur: amener de quoi écrire

TME

- Travail sur machine
- Implémentation, expérimentations et analyse
- Rendu obligatoire chaque semaine (pas noté chaque semaine, mais regardé au moment de noter le projet)

IA & DS à Sorbonne Université

Laboratoires de recherche

- **LIP6**: laboratoire d'informatique
- **ISIR**: Institut des Systèmes Intelligents et de Robotique
- d'autres laboratoires (en maths, notamment le LPSM, et des utilisateurs dans bien d'autres disciplines)

Équipes de recherche (sous-ensemble d'un laboratoire)

- **MLIA**@ISIR: Machine Learning and Information Access
- **LFI**@LIP6: Learning Fuzzy and Intelligent System
- d'autres équipes dans plusieurs laboratoires

Masters

- Master **MIND** - Intelligence artificielle, apprentissage et sciences des données
- Master **AI2D** - Algorithmes, Intelligence Artificielle, Interactions et Décision

Plus largement - SCAI

Sorbonne Center for AI: https://scai.sorbonne-universite.fr

Gens

- 150 chercheuses et chercheurs recensé.e.s autour de l'IA
- 400 doctorantes et doctorants depuis la création en 2019

Moyens

- Puissance de calcul: serveurs et GPU
- Financement: thèses, matériel

Coordination

- Projets et collaborations dans les trois facultés
- Partenaires industriels

Formations

- Plusieurs autres masters moins centrés spécifiquement sur l'IA

Période 1 – Fondations de l'apprentissage

- Naissance en même temps que les ordinateurs
- Ensemble d'outils venant des statistiques, de la géométrie, de l'optimisation et de l'algorithmique

Quelques dates

- 1950 – *Alan Turing*: test de Turing
- 1956 – *Arthur Samuel*: premier programme apprenant
- 1958 – *Frank Rosenblatt*: perceptron
- 1979 – *Kunihilo Fukushima*: réseaux de neurones modernes *Neocognitron*

Période 2 – Preuves de concept

- Premiers algorithmes pour les masses de données
- Premiers succès industriels

Quelques dates

- 1986 – *Geoffrey Hinton*: Rétro-propagation du gradient
- 1990 – *Yann Le Cun*: Lecture de chèques et de codes postaux
- 1992 – *Vladimir Vapnik*: algorithme Support Vector Machine (SVM)
- 1993 – *Ross Quinla*: algorithme arbres de décision C4.5
- 1994 – *Ernst Dickmanns*: 1000km en véhicule autonome
- 1997 – *IBM*: Deep Blue pour les échecs

Période 3 – Industrialisation

- Multiplication des outils efficaces
- Mises en production
- Disparition des réseaux de neurones

Quelques dates

- 1994 – compagnies Yahoo et Amazon
- 1997 – logiciel Weka
- 1999 – *Lee & Sung*: factorisation matricielle
- 2001 – *Lin*: bibliothèque libSVM
- 2001 – *Leo Breiman*: algorithme forêts aléatoires
- 2002 – *Pang & Lee*: classification de sentiments
- 2004 – compétition DARPA: véhicules autonomes
- 2006 – compétition Netflix: recommendation de films

Période 4 – Succès et explosion

- Traitement effication de grandes masses de données
- Réussite dans de nombreuses tâches difficiles
- Retour des réseaux de neurones

Quelques dates

- 2005 – victoire DARPA
- 2007 – langage *CUDA* pour la programmation GPU
- 2009 – logiciel *Hadoop* pour le traitement de données massivement parallèle
- 2010 – assistant vocal grand public *Siri*
- 2012 – *Alex Krivhezky* victoire ImageNet: deep learning
- 2014 – logiciel *Spark*

Période 5 – Deep learning

- Réseaux de neurones partout
- Tâches anciennes résolues et nouvelles tâches
- Bibliothèques et logicels matures

Fondations du succès

- Apprentissage: *scikit-learn*
- Réseaux de neurones: *PyTorch*, *TensorFlow*
- *Modèles de fondation* pour les images et le texte (pré-entraîné et utilisable pour diverses tâches)
- *Cloud computing*: puissance facilement disponible
- Réduction des coûts: calcul, stockage et surtout **étiquetage**

Toujours rien d'intelligent

- C'est juste un gros tas de chiffres

Intelligence artificielle et science des données

Lien très récent !

Historiquement

- IA 40: cryptographie, machine de Turing
- IA 56: tout algorithme malin
- IA 58 (a posteriori) : le perceptron de Rosenblatt
- IA 60: démonstration mathématique automatique
- IA 65: système expert
- IA 80: invention des réseaux de neurones
- IA 80: optimisation & logistique
- IA 2000: Big data
- IA 2010: Data-science
- IA 2020: intelligence artificielle explicable (xAI)
- IA 2022: chatbots & modèles génératifs

Période 6 – Modèles génératifs

ChatGPT & autres

- **Plein** d'autres: Claude, Gemini, Deepseek, et des tas et des tas

Toujours du deep learning

- Architectures nouvelles
- Calcul massivement sur GPU (cartes graphiques) pour l'entraînement et l'inférence
- Toujours plus de données
- Moins besoin d'étiquettage pour les données
- Évaluation et modèles et intégration des préférences humaines

Toujours rien d'intelligent

- Mais ça marche pour vraiment beaucoup de choses

Changement méthodologique

Modélisation

- Décrire (implémenter) dans le système les connaissances humaines du phénomène
- Algorithmique: réaliser une tâche c'est décrire les différentes étapes nécessaires
- Simulation: modéliser numériquement un phénomène physique complexe

Quelques données pour régler les degrés de libertés

Data Science

- Ensemble de données qui décrivent la tâche à accomplir, c'est à dire le résultat attendu
- Mécanisme général pour apprendre à réaliser la tâche

Beaucoup de données qui servent d'exemples

Risques de l'apprentissage

Étudiez les sorties de https://thispersondoesnotexist.com/

Qu'en pensez-vous ?

Qu'est ce qu'on apprend ?

Ce qu'il y a dans les données

- Biais de genre, de race, etc
- Rien d'objectif, au contraire
- On *reproduit* ce qui existe

Exemples (de *vrais* exemples):

- Facebook: personne noire sur une photo $\rightarrow$ singe
- CAF, FranceTravail: ciblage des contrôles des les personnes les plus fragiles
- Recrutement: jamais de personnes du 93, jamais de femmes dans l'informatique
- Chatbot: "Décris-moi une personne normale" $\rightarrow$ "Un homme blanc dans la 30aine, etc"

Solutions

Étudier les données

- Est-ce que les exemples traduisent la réalité ?
- Question simple “Y a-t-il des femmes dans mes données ?”

Meilleures méthodes d'apprentissage

- Sujets de recherche
- Par exemple: ne pas regarder le département de naissance pour un recrutement
- Difficile et peu convaincant à court terme

Problème avant tout social et politique

- Donc avant tout des solutions sociales et politiques
- Solutions informatiques non-IA: tests

Au tableau

Représentation des données

- Codage de l'information
- Robustesse
- Pré-traitements

Frontière de décision

- Exemple de la classificaiton
- Droite, cercle, formes plus compliquées
- Différentes difficultés d'apprentissage avec différents nombres de paramètres

Erreurs

- Inévitables du fait de la forme de la frontière
- Inévitables du fait de la nature des données