Master STEPE-PSL : Cours d'apprentissage automatique

Programme : Master 2 / Semestre 1

Contact : Chloé-Agathe Azencott, chloe-agathe.azencott@mines-paristech.fr

Crédits ECTS : 4

Contexte
L'essor de la numérisation fait s'accumuler des masses considérables de données et d'images, induisant un besoin croissant de fouille et d'exploitation automatisée et intelligente de ces données. De nombreux algorithmes (parmi lesquels les réseaux de neurones, les SVM, ou les forêts aléatoires) permettent des analyses et modélisations plus puissantes que la statistique classique. Le cours présente un panorama des techniques d'apprentissage automatique (ou *machine learning*), leur cadre théorique et méthodologique commun, et divers types d'applications.

Contenus
Le cours présentera une typologie du machine learning, et abordera principalement les méthodes supervisées (apprentissage prédictif), sous l'angle de la minimisation du risque empirique. Il abordera les notions de généralisation, d'évaluation et de sélection de modèle. À travers l'exemple de la régression paramétrique, le lien sera fait entre apprentissage bayésien, maximisation de la vraisemblance et minimisation du risque empirique. Le cours abordera les notions de régularisation et de méthodes à noyaux, et présentera les machines à vecteurs de support, les méthodes ensemblistes et l'apprentissage profond. Les applications de celui-ci à l'analyse d'images et aux données de la géoscience seront étudiées en détail.

Objectifs pédagogiques
- Formuler un problème d'analyse de données comme un problème d'apprentissage automatique ;
- Choisir un ou des algorithmes appropriés, en particulier pour une tâche d'apprentissage supervisé ;
- Sélectionner le modèle le plus pertinent parmi ceux générés par le ou les algorithmes choisis, en évitant le surapprentissage ;
- Évaluer la performance d'un modèle dans un scénario réaliste ;
- Mener à bout un projet d'apprentissage supervisé en Python grâce à scikit-learn.

Prérequis
Notions de probabilités et statistiques, familiarité avec Python. Des notions d'algèbre linéaire sont un plus.

Modalités d'enseignement
Le cours présente succinctement les divers paradigmes et leurs types d'applications, et laisse une place importante (environ la moitié des séances) à des Travaux Pratiques permettant une mise en œuvre concrète des principales techniques présentées.

Ce cours compte pour 4 crédits ECTS. Selon le processus de Bologne, il correspond donc à une centaine d'heures de travail. Ce travail est réparti sur 8 semaines. Vous pouvez donc prévoir une douzaine d'heures de travail sur chacune de ces semaines, soit donc en moyenne huit heures de travail hebdomadaire en plus des séances à l'emploi du temps.

Ce cours combinera des séances de cours au tableau et des travaux pratiques. Veuillez pour cela amener votre ordinateur personnel muni d'une connection Internet. Si cela est susceptible de poser problème (pas d'ordinateur, ordinateur en panne, etc.), contactez Chloé Azencott.

En dehors des heures de cours, favorisez l'utilisation du forum Moodle pour poser vos questions.

Voir la section **Programme détaillé** pour l'emploi du temps détaillé.

Ces modalités sont susceptibles d'évoluer en fonction de la situation sanitaire.

Évaluation
L'évaluation est prévue sur la base
- d'un projet numérique en Python, pour lequel il faudra rendre un court rapport ;
- d'un examen sur table avec documents autorisés.

Équipe pédagogique
Responsable de cours : Chloé-Agathe Azencott, enseignante-chercheuse au Centre de BioInformatique (CBIO) de Mines ParisTech.
contact : chloe-agathe.azencott@mines-paristech.fr / http://cazencott.info

Chargé de TP : Hugo Rollin, doctorant au Centre de Géosciences de Mines ParisTech.
contact : hugo.rollin@mines-paristech.fr

Intervenants extérieurs :
- Santiago Velasco-Forero, enseignant-chercheur au Centre de Morphologie Mathématique de Mines ParisTech.
- Nicolas Desassis, enseignant-chercheur au Centre de Géosciences de Mines ParisTech.

Ressources
Des ressources complémentaires (texte et, dans une moindre mesure, vidéo) vous sont proposées dans la section **Ressources complémentaires.**

Situation sanitaire (COVID-19)
Votre santé et celle de vos proches est votre priorité.

Si vous êtes COVID+, présentez des symptômes COVID, ou êtes cas contact d'une personne dans cette situation, **ne venez pas en cours.** Reportez-vous aux informations officielles sur https://solidarites-sante.gouv.fr/soins-et-maladies/maladies/maladies-infectieuses/coronavirus/tout-savoir-sur-la-covid-19/ ou https://www.gouvernement.fr/info-coronavirus pour plus d'information sur les symptômes et la définition des cas contact. Prévenez Chloé Azencott dès que possible.

Si vous devez suivre le cours à distance, reportez-vous aux sections du poly et vidéo complémentaires indiquées dans la section **Ressources complémentaires** pour le cours ; vous serez aussi en mesure de faire les TP par vous-même. Enfin, vous pourrez utiliser le forum Moodle pour vos questions.

Les modalités d'enseignement et le calendrier sont susceptibles d'évoluer d'une semaine sur l'autre en fonction de la situation sanitaire.

Programme détaillé

Semaine 1 (37)

Me	9/09	14h-14h50	CA Azencott	Cours : Typologie du ML, généralisation
Me	9/09	15h-15h50	CA Azencott	Cours : Sélection de modèle
Je	10/09	14h-14h50	CA Azencott	Cours : Évaluation de modèle
Je	10/09	15h-15h50	H Rollin	Notebook 1 : Sélection de modèle avec scikit-learn

Semaine 2 (38)

Me	16/09	14h-14h50	CA Azencott	Cours : Apprentissage bayésien
Me	16/09	15h-15h50	CA Azencott	Cours : Minimiation du risque empirique
Je	17/09	14h-14h50	CA Azencott	Cours : Régressions linéaires et logistiques
Je	17/09	15h-15h50	H Rollin	Notebook 2a : Régressions linéaires

Semaine 3 (41)

Me	7/10	14h-14h50	CA Azencott	Cours : Régularisation
Me	7/10	15h-15h50	H Rollin	Notebook 2b : Régressions linéaires régularisées
Je	8/10	14h-14h50	CA Azencott	Cours : SVM et noyaux
Je	8/10	15h-15h50	H Rollin	Notebook 3 : SVM et noyaux

Semaine 4 (42)

Me	14/10	14h-14h50	CA Azencott	Cours : Sélection de variables
Me	14/10	15h-15h50	H Rollin	Projet
Je	15/10	14h-14h50	CA Azencott	Cours : Réduction de dimension
Je	15/10	15h-15h50	H Rollin	Notebook 4 : Réduction de dimension

Semaine 5 (43)

Me	**21/10**	**14h-15h50**	**Évaluation à mi-parcours (avec documents)**	
Je	22/10	14h-14h50	CA Azencott	Cours : Arbres de décision, méthodes ensemblistes
Je	22/10	15h-15h50	H Rollin	Projet

Semaine 6 (44)

Me	28/10	14h-14h50	CA Azencott	Cours : Clustering
Me	28/10	15h-15h50	H Rollin	Projet
Je	29/10	14h-14h50	S Velasco-Forero	Apprentissage profond et images
Je	29/10	15h-15h50	S Velasco-Forero	Apprentissage profond et images

Semaine 7 (45)

Me	4/11	14h-14h50	S Velasco-Forero	Apprentissage profond et images
Me	4/11	15h-15h50	S Velasco-Forero	Apprentissage profond et images
Je	5/11	14h-14h50	N Desassis	Applications en géosciences
Je	5/11	15h-15h50	N Desassis	Applications en géosciences

Semaine 8 (46)

Je	12/11	14h-14h50	N Desassis	Applications en géosciences
Je	12/11	15h-15h50	N Desassis	Applications en géosciences
Ve	**13/11**	**Rendu de projet**		

Ressources complémentaires

Vous pouvez compléter le cours avec les documents suivants :

- [IML] : Introduction au Machine Learning, Chloé-Agathe Azencott, Dunod InfoSup. PDF en ligne gratuit : http://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf
- [SDD] : Polycopié du cours Science des Données, Chloé-Agathe Azencott, Mines ParisTech. https://github.com/chagaz/sdd_2020/blob/master/poly/sdd_2020_poly.pdf

Ces deux documents ont été créés à destination d'étudiant·e·s en maths appliquées. Selon vos appétances, vous en trouverez les contenus plutôt lourds en formulations mathématiques. N'hésitez pas à sauter des détails, ignorer les preuves, etc : piochez dans ce cours ce dont vous avez besoin pour vous faire une idée de ce à quoi servent les différents outils qui vous sont présentés, du contexte dans lequel ils sont applicables, de leurs points forts et de leurs inconvénients.

Des liens vers des cours en ligne OpenClassroms (formation dont Chloé-Agathe Azencott est responsable) vous sont aussi proposés plus bas. Attention, avec un compte gratuit, le nombre de vidéos que vous pouvez consulter par semaine est limité à 5.

Polycopié(s) :

- Typologie et vocabulaire : IML, Chapitre 1
- Généralisation :
  - SDD, Chapitre 8, sections 8.1.1 et 8.1.2
  - ou IML, Chapitre 2, sections 2.5.1 et 2.5.2
- Sélection de modèle : IML, Chapitre 3, section 3.1 (sauf 3.1.4)
- Évaluation de modèle : IML, Chapitre 3, section 3.2 (ne cherchez pas à retenir les différents critères qui vous sont présentés, mais plutôt à vous familiariser avec leur existence, afin d'être en mesure de revenir à cette section pour choisir un critère d'évaluation le cas échéant).
- Minimisation du risque empirique : SDD, Chapitre 7, sections 7.1 à 7.5.2
- Lien avec la maximisation de la vraisemblance : SDD, Chapitre 7, sections 7.5.3 et 7.5.4
- Régression linéaire : SDD, Chapitre 7, section 7.6
- Régression logistique : IML, Chapitre 5, section 5.3.
- Régression polynomiale :
  - IML, Chapitre 5, Section 5.4
  - ou SDD, Chapitre 9, Section 9.1.1.
- Régularisation :
  - SDD, Chapitre 8, Section 8.4
  - ou IML, Chapitre 6, Sections 6.1, 6.2 et 6.3.
- SVM linéaire : IML, Chapitre 10, sections 10.1 et 10.2
- Noyaux : IML, Chapitre 10, section 10.3
- Réduction de dimension : SDD, Chapitre 5, sections 5.3 à 5.5
- Arbres de décision : SDD, Chapitre 9, sections 9.3.1, 9.3.2 et 9.3.3
- Méthodes ensemblistes : SDD, Chapitre 9, sections 9.3.4, 9.3.5 et 9.3.6
- Clustering : IML, Chapitre 12

Textes et vidéos sur OpenClassrooms :

- Typologie et vocabulaire :
  https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/4020611-identifiez-les-differents-types-dapprentissage-automatiques
- Généralisation et sur-apprentissage :
  https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4297218-comprenez-ce-qui-fait-un-bon-modele-d-apprentissage
- Sélection de modèle :
  https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308241-mettez-en-place-un-cadre-de-validation-croisee
- Évaluation de modèle :
  - Classification (prédiction binaire) :
    https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308256-evaluez-un-algorithme-de-classification-qui-retourne-des-valeurs-binaires
  - Classification (prédiction continue) :
    https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308261-evaluez-un-algorithme-de-classification-qui-retourne-des-scores
  - Régression :
    https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308276-evaluez-un-algorithme-de-regression
- Régression linéaire :
  https://openclassrooms.com/fr/courses/4444646-entrainez-un-modele-predictif-lineaire/4444653-trouvez-une-combinaison-lineaire-de-variables-qui-approxime-leurs-etiquettes
- Régression logistique :
  https://openclassrooms.com/fr/courses/4444646-entrainez-un-modele-predictif-lineaire/4507831-predisez-lineairement-la-probabilite-de-l-appartenance-d-un-point-a-une-classe
- Régularisation :
  - Concept de régularisation :
    https://openclassrooms.com/fr/courses/4444646-entrainez-un-modele-predictif-lineaire/4507791-controlez-la-complexite-de-votre-modele
  - Régression ridge :
    https://openclassrooms.com/fr/courses/4444646-entrainez-un-modele-predictif-lineaire/4507801-reduisez-l-amplitude-des-poids-affectes-a-vos-variables
- Régression lasso :
  https://openclassrooms.com/fr/courses/4444646-entrainez-un-modele-predictif-lineaire/4507806-reduisez-le-nombre-de-variables-utilisees-par-votre-modele
- SVM linéaire :
  https://openclassrooms.com/fr/courses/4444646-entrainez-un-modele-predictif-lineaire/4507841-maximisez-la-marge-de-separation-entre-vos-classes
- SVM à noyau :
  https://openclassrooms.com/fr/courses/4470406-utilisez-des-modeles-supervises-non-lineaires/4722466-classifiez-vos-donnees-avec-une-svm-a-noyau
- Réduction de dimension :
  https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises/4379443-comprenez-pourquoi-reduire-la-dimension-de-vos-donnees

- Analyse en composantes principales : https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises/4379481-calculez-les-composantes-principales-de-vos-donnees
- Arbres de décision et méthodes ensemblistes : https://openclassrooms.com/fr/courses/4470521-modelisez-vos-donnees-avec-les-methodes-ensemblistes
- Clustering : https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises (Partie 3)