Apprentissage Statistique
STT 3790

Automne 2024 — 3 crédits

professeur : Florian Maire — 4253, Pavillon André Aisenstadt
courriel : florian.maire@umontreal.ca

Horaire :
- Les cours magistraux auront lieu au local **1177, A. Aisenstadt**, les
  - mardi de 15h30 à 17h30
  - vendredi de 8h30 à 9h30

  Premier cours le mardi 3 septembre. Dernier cours le vendredi 6 décembre.
- Les séances de TP se dérouleront les vendredi de 12h30 à 14h30 au local **1177, A. Aisenstadt**. Premier TP le vendredi 13 septembre, dernier TP le vendredi 6 décembre.
- Pas de cours/TP la semaine du 21 octobre (relâche).

Auxiliaire d'enseignement : Guillaume Villeneuve, guillaume.villeneuve.2@umontreal.ca.

Présentation : L'apprentissage statistique est une discipline qui concerne l'extraction d'informations à partir de données observées en vue de résoudre un problème précis. Il peut s'agir par exemple de détecter des motifs dans des signaux (images, séries chronologiques), d'identifier des liens d'association entre différentes variables ou encore de mettre en évidence des structures de groupe dans un jeu de données. D'une façon plus abstraite, ceci revient à trois familles de problèmes connues en statistique en tant que modélisation, régression et classification. Les deux éléments de départ de ce cours sont donc (i) le problème à résoudre et (ii) la nature des données. Ce cours se concentre sur la présentation et l'étude d'algorithmes. Un algorithme consiste en une suite d'instructions programmable sur un ordinateur. Outre des considérations d'ordre conceptuel, théorique et computationnel, les algorithmes seront illustrés sur ces différents problèmes rencontrés en statistiques.

Pour tirer un maximum de profit de ce cours (et le réussir), il faut que les notions vues dans le cours de Régression linéaire (STT2400) et Concepts et méthodes en statistique (STT2700) soient très bien assimilées. Des résultats du cours de Probabilités (MAT1720) et l'Algèbre linéaire (MAT1600) seront aussi fréquemment utilisés. Ce cours est relié à deux cours de troisième année : Modélisation prédictive (STT3261) qui étudie certains problèmes vus dans STT3790 de manières plus précises et Fondements théoriques en science des données (STT3795).

Fonctionnement du cours : Les notes de cours (pour chaque chapitre) sont mises en ligne sur Studium, au fur et à mesure. Les cours magistraux se font généralement au tableau : on travaille aussi bien sur les éléments des notes (définitions, propriétés, preuves, exercices, etc.) que sur des exemples ou des digressions qui ne sont pas dans les notes. Occasionnellement, des illustrations sur ordinateur (simulations, code d'algorithmes, etc.) seront présentées en classe. Les feuilles d'exercices de TP sont mises en ligne à l'avance. Elles contiennent aussi bien des questions théoriques que des questions de programmation (le language de programmation est toujours laissé libre dans ce cours). Il est nécessaire de travailler sur ces feuilles de TP dès qu'elles sont en ligne (il est recommendé de travailler en petits groupes) pour que la séance de TP soit intéressante et fructueuse.

Disponibilités : Quatre heures de disponibilités seront offertes chaque semaine,
- Florian Maire : mon bureau (AA 4253) chaque lundi de 13h à 14h et mardi de 13h à 14h.
- Guillaume Villeneuve : centre aide, bibliothèque maths-info, chaque jeudi 13h30 à 15h30.

Description détaillée :

Introduction

- Apprentissage statistique et apprentissage machine ;
- Apprentissage supervisé ;
- Apprentissage non supervisé ;

Partie 1 : Apprentissage supervisé

Chapitre 1 Régression linéaire
- Modèle de régression linéaire ;
- Estimateur et fonction de perte et minimisation du risque d'apprentissage ;
- Algorithme *Iteratively Reweighted Least Squares* ;

Chapitre 2 Réduction de dimension
- Concept d'information ;
- Analyse en composantes principales ;
- Régression sur composantes principales ;

Chapitre 3 Méthodes de rétrécissement
- Compromis biais-variance ;
- Estimateurs *Ridge* et *Lasso* ;
- Choix du facteur de rétrécissement par validation croisée ;

Chapitre 4 Modèles linéaires généralisés
- Famille exponentielle et fonction de liens ;
- Régression : algorithmes du gradient, de Newton-Raphson et de Fisher ;
- Déviance et sélection de modèles ;

Chapitre 5 Classification paramètrique
- Régression logistique ;
- Perceptron ;
- Classifieur naïf de Bayes ;

Partie 2 : Apprentissage non supervisé

Chapitre 6 Méthodes de regroupement (*clustering*)
- Algorithme des k-moyennes (*k-means clustering*) ;
- Modèle paramétrique : algorithme *Expectation-Maximization* et échantillonneur de Gibbs ;
- Modèle non-paramétrique : processus de Dirichlet.

Des notes de cours concernant ces différents chapitres seront mises en ligne sur Studium au fur et à mesure du cours.

Évaluation du cours :

- **Quiz ou QCM** Des questions courtes avec ou sans développement à faire individuellement d'une durée de 15-20 minutes.
- **Devoirs** Les devoirs se feront en équipe de 4 ou 5 étudiant-es. Les devoirs sont postés sur studium environ deux semaines avant la date de remise (voir ci-dessous). Le format des devoirs sera varié suivant la partie du cours : question théorique et/ou mini-projet d'analyse de jeu de données ou de simulation.
- **Intra & Final** Ces examens évaluent les capacités individuelles et se dérouleront en salle d'examen (pas de projet ou take-home). Toute absence lors de ces examens devra être justifiée et il appartiendra à l'autorité compétente de déterminer si le motif est acceptable.

type	date	pondération
Quiz 1	en TP le 20 septembre	5%
Quiz 2	en TP le 4 octobre	5%
Devoir 1	à rendre le 18 octobre	10%
Devoir 2	à rendre le 22 novembre	10%
Intra	le 1 novembre de 12h30 à 14h30 (A. Aisenstadt, 1177)	30%
Final	le 13 décembre de 12h30 à 15h30 (A. Aisenstadt, 1177)	40%

Les étudiants inscrits au Bureau de Soutien aux Étudiants en Situation de Handicap (ESH) désirant bénéficier de mesures d'accommodement aux examens (intra et final) sont priés de cliquer ici pour connaître la procédure à suivre.

Plagiat : L'Université de Montréal a une politique très claire sur le plagiat que vous êtes invités à consulter au `www.integrite.umontreal.ca`. Elle ne concerne pas que les examens, mais également les devoirs.

Bibliographie : Le programme de ce cours n'est pas exactement calqué sur le développement d'un livre en particulier. Pour la Partie 1 du cours (apprentissage supervisé), on pourra se référer aux trois livres suivants.

- Efron, B. et Hastie, T. (2021). *Computer Age Statistical Inference*, Cambridge University Press.
  => on pourra lire les chapitres 1–8, 16 et 17
- James, G., Witten, D., Hastie, T. et Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*, Springer
  => on pourra lire les chapitres 1–5, 8 et 10
- Hastie, D., Tibshirani, R. et Friedman, J. (2017). *The Elements of Statistical Learning*, Springer.
  => on pourra lire les chapitres 1–4, 7, 9, 10 et 15
- Dobson, A. et Barnett, A. (2008) *An Introduction to Generalized Linear Models*, CRC Press.
  => on pourra lire les chapitres 3–5, 7 et 8

Pour la Partie 2 du cours (apprentissage non supervisé), on pourra se référer à

- Efron, B. et Hastie, T. (2021). *Computer Age Statistical Inference*, Cambridge University Press.
  => on pourra lire les chapitres 9
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer.
  => on pourra lire les chapitres 9 et 11

Ces livres sont accessibles au format numérique sur le site de la bibliothèque de l'udem.

Par rapport au Programme d'agrément universitaire (PAU) de l'Institut canadien des actuaires (ICA) :

Le cours STT3790 fait partie de la liste des cours obligatoires pour la reconnaissance du diplôme de baccalauréat auprès de l'ICA. À noter que :

1. Ce cours est agréé en vertu du Programme d'agrément universitaire (PAU) de l'Institut canadien des actuaires (ICA) pour l'année universitaire 2023-2024. Ce cours fait partie des cours nécessaires à réussir pour le crédit du diplôme menant à l'admission pour l'examen synthèse de l'ICA du titre AICA. Veuillez consulter les pages suivantes pour de plus amples détails :

   Programme de crédit par diplôme

   https ://education.cia-ica.ca/fr/accueil/

   https ://education.cia-ica.ca/fr/universites/

2. En plus des politiques internes en matière de comportements spécifiques à une université, y compris l'inconduite universitaire, les candidats désirant obtenir des crédits aux examens professionnels seront également assujettis à la Politique relative au Code de conduite et d'éthique des candidats faisant partie du système d'éducation de l'ICA ainsi qu'au Code de conduite et d'éthique pour les candidats au titre d'actuaire dans le système d'éducation de l'ICA :

   https ://www.cia-ica.ca/docs/default-source/2020/220064f.pdf

   https ://www.cia-ica.ca/docs/default-source/2020/220065f.pdf