Apprentissage Statistique
STT 3790

Automne 2020 — 3 crédits
Professeur : Florian Maire — 4243, Pavillon André-Aisenstadt
Tél. : 343-7977 — maire@dms.umontreal.ca

Horaire : Les cours magistraux auront lieu les mardi de 13h30 à 14h30 et vendredi de 10h30 à 12h30 du 1er septembre jusqu'au 8 décembre (sauf les 20 et 23 octobre). Les cours se feront au tableau et seront retransmis en direct sur zoom, le lien pour chaque cours est sur studium (un lien pour les mardi et un lien pour les vendredi). Étant donné l'absence d'auxiliaire, il n'y a pas de séance de mise en pratique (initialement prévues les mardi de 10h30 à 12h30). Je donnerai à la place des exercices, des analyses de jeux de données ou des simulations à faire dont des solutionnaires seront mises en ligne réguliérement. Certains devoirs seront évalués (voir section évaluation). Cependant, les deux heures du mardi (10h30-12h30) pourront être utilisées, par exemple pour faire cours.

Concernant l'agrément de l'ICA :

- Ce cours est agréé en vertu du Programme d'agrément universitaire (PAU) de l'Institut canadien des actuaires (ICA) pour l'année universitaire 2020-2021. L'atteinte de la note minimale établie pour ce cours peut permettre à un candidat d'obtenir un crédit de l'ICA pour certains examens d'actuariat préliminaires. La note minimale pour ce cours est B. En raison du contexte de cours en ligne liés à la COVID-19, la note minimale pourrait être révisée à la hausse suite à la vérification qui aura lieu lorsque la session sera complétée. Veuillez noter qu'une combinaison de cours pourrait être nécessaire pour obtenir un seul crédit d'examen. Veuillez consulter la page suivante pour de plus amples détails : `https://www.cia-ica.ca/fr/adhesion/programme-d-agrément-universitaire-(pau)-de-l-ica`.
- 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 de formation de l'ICA : `https://www.cia-ica.ca/docs/default-source/2016/216106f.pdf` et `http://www.cia-ica.ca/docs/default-source/2016/216107f.pdf`.

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, consistant d'une suite d'instructions programmable sur un ordinateur, permettant de faire une analyse statistique minutieuse de ces différentes situations. Outre des considérations d'ordre conceptuel, théorique et computationnel, les algorithmes seront illustrées sur ces différents problèmes rencontrés en statistiques. En apprentissage supervisé, on connaît un certain nombre d'informations additionnelles : c'est le cas par exemple des problèmes de régression où l'on connaît des variables explicatives. Tout comme en apprentissage machine, on distingue en apprentissage statistique deux types de situations limites : l'apprentissage supervisé et non-supervisé. Ces deux cadres se différencient à plusieurs niveaux et ce cours les traitera en deux parties disctinctes.

Description détaillée :

Partie I Apprentissage supervisé

Chapitre 1 Modèle linéaire et régression
- moindres carrés généralisés
- algorithme *Feasible Generalized Least Squares*
- régression sur composantes principales

Chapitre 2 Méthodes de rétrecissement
- compromis biais-variance
- estimateur de James-Stein et Bayes empirique
- régressions Ridge et Lasso

Chapitre 3 Modèles linéaires généralisés
- famille exponentielle et fonction de liens
- régression : algorithmes du gradient, de Newton-Raphson, de Fisher
- déviance et sélection de modèles

Chapitre 4 Classification
- régression logistique (et perceptron multicouche)
- algorithme des K plus proches voisins (et sa version probabiliste)
- classifieur naïf de Bayes

Chapitre 5 Méthodes par arbres
- arbre de régression et arbre de classification
- forêt aléatoire (*bagging* et *boosting*)

Partie II Apprentissage non supervisé

Chapitre 1 Méthodes de regroupement (*clustering*)
- Algorithme des K moyennes
- Regroupement hiérarchique

Chapitre 2 Modèles à variables latentes
- Regroupement probabiliste
- Modèle de mélange Gaussien
- Algorithme EM

Chapitre 3 Approche Bayésienne
- Modèle hiérarchique : mélange fini et infini (processus de Dirichlet)
- Échantillonneur de Gibbs

Disponibilités : Deux heures de disponibilités seront offertes chaque semaine, elles auront lieues sur zoom, chaque mercredi de 15h30 à 17h30.

Évaluation du cours :

- **Devoirs** À partir de la deuxième semaine du cours il y aura un devoir toutes les deux semaines. Généralement les devoirs se feront en équipe mais, pour certains devoirs, ils seront à faire obligatoirement de manière individuelle. Ils consistent en une question reliée à la matière vue précédemment en cours. Les devoirs sont postés sur studium le vendredi (à partir du 11 septembre) et sont à remettre avant la séance de mise en pratique le mardi suivant (donc le premier devoir sera à remettre le 15 septembre). Le format de ces questions sera varié suivant la partie du cours : question théorique ou mini-projet d'analyse de jeu de données ou de simulation.
- **Intra & Final** Ce sont des examens individuels avec questions personalisées à développement qui se déroulement en ligne dans des circonstances strictes en lien avec les exigences de l'ICA. L'intra aura lieu en le vendredi 16 octobre et le final le 4 décembre, tous les deux sur la plage horaire du cours.

type	date	pondération
Devoirs	–	20%
Devoirs individuels	–	5%
Intra	16 octobre	35%
Final	4 décembre	40%

Les étudiants inscrits au Bureau de Soutien aux Étudiants en Situation de Handicap (BSESH) désirant bénéficier de mesures d'accommodement aux examens (intra et final) sont priés de consulter le lien suivant pour connaitre la procédure à suivre : `https://safire.umontreal.ca/reussite-et-ressources/mesures-daccommodement-aux-examens-pour-les-etudiants-en-situation-de-handicap/`.

Autres règles : La date limite pour abandonner le cours « sans frais » est le 17 septembre alors « qu'avec frais » (sans être remboursé), c'est le 6 novembre. Par la suite, si vous abandonnez, vous aurez un échec. Toute absence lors des examens en ligne devra être obligatoirement justifiée et il appartiendra à l'autorité compétente de déterminer si le motif est acceptable.

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 I du cours (apprentissage supervisé), on pourra se réferrer aux trois livres suivants.

- 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 II du cours (apprentissage non-supervisé), on pourra se réferrer à

- 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.