STT 6415–Régression

Hiver 2024
Professeur : Florian Maire
florian.maire@umontreal.ca
bureau : A. Aisenstadt, 4253.

Horaires et déroulement du cours

— les mardi de 14h30 à 16h00, local 5183, AA,
— les jeudi de 14h30 à 16h00, local 5183, AA,
du mardi 9 janvier jusqu'au jeudi 25 avril (sauf pour la semaine de lecture). Les cours se feront essentiellement au tableau et s'appuieront sur des notes de cours qui seront fournies à l'avance. Il y aura une période de questions offerte aux étudiants qui se déroulera chaque semaine de cours
— les mercredi de 10h00 à 12h00, local 4253, AA.

Présentation

Ce cours porte sur certains aspects avancés de l'étude des modèles de régression, utilisés en statistique et apprentissage automatique. Un modèle de régression établi une relation du type $Y = f(X) + \varepsilon$ où Y, X, ε sont des variables aléatoires (réponse, covariable, bruit) et $f : \mathbb{R}^d \to \mathbb{R}$ est la fonction (déterministe) de régression. En apprentissage statistique et automatique, le travail se concentre sur l'estimation de f ayant observé des données $\{(X_i, Y_i) : i \leq n\}$. Après une analyse du cas du modèle de régression linéaire qui se concentrera sur la situation où d peut être plus grand que n, ce cours traîtera essentiellement de régression non-paramètrique. De ce points de vue, l'articulation se fera en deux temps : tout d'abord, nous adopterons le point de vue fréquentiste pour étudier certaines techniques de régression telles que celles basées sur des noyaux et des partitions avant d'en arriver aux estimateurs à splines. Dans un second temps, nous adopterons le point de vue bayésien. En particulier, nous travaillerons sur les estimateurs bayésiens basés sur une loi a priori dont le support est l'ensemble des processus gaussiens. Enfin, nous concluerons en établissant une correspondance entre les deux approches.

Description détaillée

— Modèle de régression linéaire : révision de l'estimateur des moindres carrés ordinaires (MCO) et quelques résultats. Analyse dans le scénario n libre et d fixe puis pour d qui grandit avec n : nécessité d'introduire des contraintes de sparsité pour résoudre les problèmes en grande dimension. Définition du risque minimax et analyse de certains résultats récents sur l'efficacité minimax des moindres carrés contraints/pénalisés dans le cas du modèle linéaire. Puis, modèle de régression quelconque et estimateur linéaire : étude de l'impact de la misspecification sur la convergence et présentation de quelques inégalités d'oracles.
— Estimateur non-paramètrique : pour commencer, nous établirons deux résultats importants sur la borne inférieure du risque minimax en régression non-paramétrique et le théorème de Stones. Ce dernier nous servira pour établir la convergence d'estimateurs locaux (estimateur à noyaux, à partition, k-NN). Ensuite, nous nous intéresserons aux estimateurs non-paramètriques définis implicitement comme la solution d'un problème d'optimisation tel que le problème des moindres carrés. L'intérêt est que ces estimateurs peuvent être choisis comme étant des fonctions lisses ou ayant certaines régularités ce qui est à la fois désirable en pratique et du point de vue de l'estimation. On étudiera particulièrement le cas des polynômes par morceaux (splines) et des estimateurs à noyau reproduisant (RKHS).
— Estimateur non-paramètrique bayésien : régression par processus gaussiens. Nous commencerons par établir des notions nécessaires pour comprendre comment faire de l'apprentissage statistique dans un cadre bayésien ainsi que par quelques rappels sur certains processus stochastiques tels que les processus gaussiens. Nous présenterons l'estimateur bayésien de la fonction de régression basé sur une loi a priori supportant les processus gaussiens puis nous étudierons quelques défis d'ordre computationnel liés à l'inférence bayésienne. En particulier, nous tâcherons de faire des liens entre cet estimateur bayésien et les autres estimateurs non-paramétriques vus précédemment. Si le temps le permet, des résultats de convergence relatifs à cet estimateur seront donnés.

Évaluations

type	date	pondération
Quiz 1	23 janvier	5%
Devoir 1	13 février	10%
Intra	27 février 14h30–16h00	30%
Quiz 2	14 mars	5%
Devoir 2	28 mars	10%
Final	à organiser (prévoir 3h)	40%

Les quizzes auront une durée de 10-20 minutes au début des cours en question, il s'agira d'une petite question ou d'un QCM. Pour les devoirs, ils seront postés sur studium environ deux semaines avant le début du cours. Enfin, 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 contacter le SAFIRE.

Plagiat

L'Université de Montréal a une politique très claire et ferme sur le plagiat, voir `https://integrite.umontreal.ca`. Elle ne concerne pas que les examens, mais également les devoirs. Ce rappel est d'autant plus valable car, par nature, l'environnement dans lequel les examens en ligne se déroulent est plus difficilement controllable. Plutôt que d'opter pour une méthode de surveillance disproportionnée, l'utilisation de toutes les ressources (livres, notes de cours, internet, logiciels) est permise lors des examens. En revanche, la communication entre étudiants est strictement interdite. À ce niveau, il sera demandé à ce que chaque étudiant écrive une déclaration sur l'honneur en introduction de leur copie d'examen, guarantissant le caractère personnel de leur travail. Il en va de la valeur de vos diplômes !

Bibliographie

Le cours ne suivra pas un manuel de référence mais il pourra être intéressant de consulter les livres suivants :

Györfi, Kohler, Krzyżak and Walk, *A distribution free theory of nonparametric regression*, 2002, Springer
Tsybakov, *Introduction to nonparametric estimation*, 2009, Springer
Rasmussen and Williams, *Gaussian processes for machine learning*, 2006, MIT
Hastie, Tibshirani and Friedman, *The elements of statistical learning*, 2009, Springer
Wainwright, *High-dimensional statistics : a non-asymptotic viewpoint*, 2019, Cambridge