Sujets pour les projets - Probabilités et Statistiques

Année universitaire 2023/2024

Sujet 1 – Simulation de variables aléatoires 1

On s'intéressera aux principes et méthodes classiques (méthode de la fonction inverse, méthode de rejet, ...) permettant la simulation de variables aléatoires usuelles à partir de variables aléatoires uniformes sur $[0,1]$ que l'on présentera rigoureusement et illustrera sur des exemples au travers de la rédaction et de l'exécution de programmes/algorithmes dans un langage à choisir.
Référence possible : *Non-Uniform Random Variate Generation*, L. Devroye 1986), Chapitre 2, `http://www.nrbook.com/devroye/`.

Sujet 2 – Simulation de variables aléatoires 2

On s'intéressera aux méthodes et algorithmes spécialisés méthode des séries, méthode de Forsythe-von Neumann, ...) permettant la simulation de variables aléatoires lorsque les méthodes classiques (du Sujet 1) sont en défaut, par exemple lorsque la densité de la variable aléatoire est donnée par un série de fonction. On présentera et étudiera rigoureusement ces méthode et on les illustrera sur des exemples au travers de la rédaction et de l'exécution de programmes/algorithmes dans un langage à choisir.
Référence possible : *Non-Uniform Random Variate Generation*, L. Devroye (1986), Chapitre 4, `http://www.nrbook.com/devroye/`.

Sujet 3 – Introduction à l'apprentissage statistique 1 : régression linéaire

La théorie de l'apprentissage statistique consiste en la recherche d'une fonction prédictive basée sur des données. Lorsque l'on cherche à expliquer une réponse numérique Y grâce à des variables $X_1, \dots, X_n$ à des fins d'inférence ou de prédiction, un des modèles les plus simple est la régression linéaire. On présentera les concepts et résultats principaux de cette théorie et on les illustrera en travaillant avec des jeux de données conséquents sous R.
Référence possible : *An Introduction to Statistical Learning with Applications in R*, G. James, D. Witten, R. Tibshirani et T. Hastie (2013), Chapitre 3 (prérequis Chapitre 2), `https://www.statlearning.com/`

Sujet 4 – Introduction à l'apprentissage statistique 2 : classification

La théorie de l'apprentissage statistique consiste en la recherche d'une fonction prédictive basée sur des données. Lorsque l'on cherche à expliquer une réponse qualitative (et non quantitative) Y grâce à des variables $X_1, \dots, X_n$ à des fins d'inférence ou de prédiction, l'idée de la régression linéaire est, évidement, mise en défaut et le concept de classification permet de répondre à ce problème (régression logistique, analyse linéaire de discriminants,...). On présentera les concepts et résultats principaux de cette théorie et on les illustrera en travaillant avec des jeux de données conséquent sous R.
Référence possible : *An Introduction to Statistical Learning with Applications in R*, G. James, D. Witten, R. Tibshirani et T. Hastie (2013), Chapitre 4 (prérequis Chapitre 2), `https://www.statlearning.com/`

Sujet 5 – Processus de Poisson et d'assurance

À l'interface entre économie et mathématiques, la théorie du risques permet de fournir des modèles permettant d'estimer les dédommagements qu'un assureur aura à verser à ses clients, prenant en compte les temps où surviennent les demandes de versement de dédommagement provoqués par des sinistres et leurs importances. Les demandes dédommagements étant espérés relativement rares et indépendants les un des autres, il est raisonnable de voir apparaître des processus de Poisson pour régir « les temps aléatoires de sinistres » $t_1 < t_2, \dots$; par ailleurs, chaque sinistre conduisant au versement d'un indemnité aléatoire d'un montant X_i (i.i.d.), il est naturel d'étudier le montant réclamé total $\sum_{i=1}^{N(t)} X_i$, où $N(t)$ est le nombre de sinistres entre les temps 0 et t. On présentera ces modèles et concepts et on illustrera les propos par des simulations numériques. En guise de prolongement, on pourra s'intéresser aux probabilités de ruine d'une compagnie d'assurance.

Référence possible : *Non-Life Insurance Mathematics*, T. Mickosch (2009), Chapitre 1 (ouverture Chapitre 4), disponible à la BU en ligne.

Sujet 6 – Introduction à la théorie des valeurs extrêmes

Étant donnée une suite de variables aléatoires réelles $(X_k)_{k\in\mathbf{N}^*}$, la Théorie des Valeurs Extrêmes (EVT) vise à l'étude du comportement (asymptotique) de $M_n = \max_{1\leq k\leq n} X_k$ et $m_n = \min_{1\leq k\leq n} X_k$ ou plus généralement aux statistiques d'ordre, dans un premier temps en supposant les X_k indépendantes puis en cherchant à relâcher cette hypothèse. Ce mémoire se veut être une introduction à cette théorie.

On pourra, par exemple, s'intéresser au Théorème de Fisher–Tippett–Gnedenko, montrant que si les X_k sont i.i.d. et s'il existe des suites de constantes de normalisation $(a_n)_{n\in\mathbf{N}^*}$ et $(b_n)_{n\in\mathbf{N}^*}$ telles que $b_n^{-1}(M_n - a_n)$ converge vers une v.a. non dégénérée celle-ci suit nécessairement une des lois max-stables (Fréchet, Gumbel ou Weibull), aux méthodes d'étude de M_n *via* le processus ponctuel des excédants ou encore à l'approche Peak-Over-the-Threshold (POT).

Référence possible : *Modeling Extremal Events for Insurance and Finance*, P. Embrechts, C. Klüppelberg et T. Mikosch, Stochastic Modelling and Applied Probability, 33, Springer Science & Business Media, 2013.

Sujet 7 –Marches aléatoires et réseaux électriques

La théorie des réseaux électriques fournit des outils intuitifs et pratiques pour étudier les chaînes de Markov réversibles sur des graphes connexes. Après une introduction à cette théorie et avoir compris les liens qu'elle entretient avec l'étude des marches au hasard, on pourra notamment en faire usage pour démontrer le théorème de Pólya sur la récurrence et la transience des marches aléatoires simples au plus proche voisin sur la grille $\mathbf{Z}^d$, $d \in \mathbf{N}^*$ et s'intéresser au questions de temps d'atteinte ou de couverture d'un sous-ensembles de sommets du réseau sous-jacent. Outre la simulation de trajectoires de de marches aléatoires, l'outil informatique pourra permettre d'illustrer certains résultats de la théorie dévoilée (mesure invariante dans le cas récurrent positif, identité pour le temps d'atteinte, ...).

Références possibles :

1. *Probability on trees and networks*, R. Lyons et Y. Peres (2016), Chapitre 2, `https://rdlyons.pages.iu.edu/prbtree/book_online.pdf`
2. *Random walks and electrical networks*, P. G. Doyle et J. L. Snell (1984), `https://arxiv.org/pdf/math/0001057.pdf`

Sujet 8 –Modèle épidémie SIR en épidémiologie : présentation et illustration

Les modèles épidémiologiques habituellement étudiés sont des modèles déterministes. Toutefois s'ils peuvent être plus difficiles à étudier et moins bien refléter la réalité, les modèles probabilistes sont une façon naturelle de modéliser l'évolution d'une épidémie : chaque individu à une certaine probabilité d'être infecté par la maladie. Une part importante de l'étude de ces problèmes stochastiques va être de déterminer si, quand la taille de la population augmente, ils convergent vers un problème déterministe.

Référence possible :

1. *Modélisation stochastique d'une épidémie SIR*, Manon Costa (2011),`https://www.math.ens.psl.eu/shared-files/10744/?introduction%20au%20domaine%20de%20recherche.pdf`

Sujet 9 –Bootstrap : compréhension et utilisation

L'estimation de la variance est un problème difficile dans les enquêtes. Les poids finaux utilisés à l'étape de l'estimation comprennent plusieurs traitements statistiques, notamment la correction de la non-réponse totale et le calage, dont l'effet sur la variance doit être évalué. Le bootstrap est un instrument utile, qui permet de créer les poids dits bootstrap publiés avec l'ensemble de données de l'enquête. Ces poids peuvent servir à calculer de façon répétée la version bootstrap du paramètre d'intérêt, ce qui donne un estimateur de la variance ou un intervalle de confiance basés sur des simulations.

Référence possible :

1. *Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages Principes, exemples et mise en œuvre*, Pascal Bessonneau (1) , Gwennaëlle Brilhaut (1) , Guillaume Chauvet (2) , Cédric Garcia (4) (2022) , `https://dumas.ccsd.cnrs.fr/INED/hal-03524669v1`

Sujet 10 –Évaluer la performance des modèles de classification de Machine Learning

Les courbes ROC (fonctions d'efficacité du récepteur) sont un outil important pour évaluer les performances d'un modèle de Machine Learning. Elles sont le plus souvent utilisées pour des problèmes de classification binaire dont la sortie est composée de deux classes distinctes. La courbe ROC montre le rapport

entre le taux de vrais positifs (TPR) du modèle et le taux de faux positifs (FPR). Le TPR est le taux auquel le classificateur prédit un résultat « positif » pour des observations qui sont « positives ». Le FPR est le taux auquel le classificateur prédit un résultat « positif » pour les observations qui sont en fait « négatives ». Un classificateur parfait aura un TPR de 1 et un FPR de 0.
Référence possible :

1. *Courbe ROC*, Xavier Dupre, `http://www.xavierdupre.fr/app/mlstatpy/helpsphinx/c_metric/roc.html#exemple`

Sujet 11 – Le modèle de Percolation

Immergeons une pierre spongieuse dans l'eau. La pierre est percée de petits canaux et l'eau commence à s'infiltrer à l'intérieur. La question est la suivante : le centre de la pierre sera-t-il mouillé ? Ce modèle de probabilités discrètes est devenu un domaine de recherche à part entière parce qu'il contient une transition de phase. Le paramètre microscopique qui définit la porosité de la pierre détermine des phénomènes macroscopiques très différent quand il dépasse une certaine probabilité de transition.
Référence possible :

1. *Percolation* , Geoffrey Grimmett.

Sujet 12 – Le collectionneur de vignettes

Vous faites la collection de vignettes présentes dans vos paquets de céréales. La collection entière comprend N vignettes différentes et vous achetez un paquet de céréales par semaine, combien de temps vous faudra-t-il pour compléter votre collection ? Après avoir travaillé sur les résultats théoriques, nous pourrons estimer ce temps dans certains cas plus généraux : vous vous alliez avec votre voisin qui fait la même collection que vous et partagez vos doublons, le fabricant de céréales vous permet d'obtenir la vignette de votre choix en échange de 10 autres vignettes, ...
Référence possible :

1. *Probability vol 1* , William Feller.

Sujet 13 – Algorithmes génétiques

On modélise des individus par leur matériel génétique : une suite finie de lettres A, T, G et C. Une population d'individu est constituée d'un nombre fini de telles suites. On étudie l'évolution d'une telle population au cours du temps. Lorsqu'un individu se reproduit, il engendre une copie de son génôme, avec quelques erreurs aléatoires. Il existe aussi une unique chaîne de A, T, G et C qui possède un avantage sélectif : les individus avec cette chaîne se reproduisent plus que les autres. Il existe ainsi deux forces antagonistes : l'une qui tend à faire naître des individu au matériel génétique unique et l'autre qui veut conserver la meilleure chaîne possible. Cette lutte se traduit dans le modèle par une transition de phase, que nous pourrons observer sur des simulations.
Référence possible :

1. *A basic model of mutations* , Berger M., Cerf R., `https://arxiv.org/abs/1806.01212`