Partiel PSB : Mathématiques pour le Big Data avec Mr.LAUDE

Thuy AUFRERE

31/01/2021

1. Introduction

Les 5 travaux mathématiques sélectionnés sont :

- **Interpolation des données spatiales** d'Arnaud Bruel YANKO, Adrien JUPYTER
- **Systèmes de conduite autonome** d'Akram BENSALEM
- **Algorithme Génétique** de Ramya WURAOLA
- **Apprentissage par Arbres de Décisions** de Corentin BRETONNIERE, Antoine SERREAU, Benjamin GUIGON
- **Time domaine approache** de Siva CHANEMOUGAM

Pourquoi j'ai sélectionné ces 5 travaux :

J'ai sélectionné ces 5 papiers car ce sont ceux dont les sujets m'intéressaient et/ou m'intriguaient et dont j'aimerais en apprendre plus.

Pour chaque travail, je commenterais selon le plan suivant :

- les informations sur le travail (nom du package, nom de l'auteur et le lien vers Github)
- une rapide synthèse reflétant ma compréhension du travail
- une explication de certaines formules mathématiques
- une évaluation du travail selon mes critères qui sont : i) sur la forme (structure, clarté, fautes d'orthographe) et ii) sur le fond (explication, exemple, données). Je n'attribuerais pas de notes pour ces travaux. En effet, n'ayant pas fait d'études de mathématiques, je ne me sens pas légitime d'attribuer une note. Cette évaluation sera suivie d'une petite conclusion concernant l'appréciation du travail.

1. Papier 1

Nom du papier : **Interpolation des données spatiales**

Auteur du travail : Arnaud Bruel YANKO, Adrien JUPYTER

Lien du travail sur Github :

- https://github.com/ARNAUD-BRUEL-YANKO/PSBX
- https://github.com/akjupiter/PSBX/tree/master/Maths

1.1. Synthèse du papier 1

Ce papier a pour sujet l'interpolation des données spatiales en prenant les deux approches qui sont le déterminisme (i) et la géostatistique (ii).

(i)- Le premier prédit des valeurs où il n'y avait pas d'échantillons disponibles, son but n'est pas d'évaluer la structure spatiale des données. Pour faire une estimation représentative de la moyenne globale, le papier indique qu'il est possible d'utiliser la méthode des polygones d'influences de Thiessen, la méthode des cellules et la géostatistique intrinsèque.

(ii)- Le deuxième permet d'ajuster des modèles spatiaux à des données. Pour calculer l'estimation globale, le papier indique qu'il est possible de calculer la variance qui représente la mesure de la précision de l'estimation. Celle-ci suppose de faire des hypothèses sur l'échantillonnage ponctuel aléatoire pur et sur l'échantillonnage non ponctuel.

Outre l'estimation globale, ce papier développe l'existence de l'estimation locale avec la technique du Krigeage (rechercher parmi les estimateurs linéaires, ceux qui ont les meilleures propriétés) et le Cokrigeage (rechercher la variabilité ou la corrélation spatiale d'une variable 1 avec l'aide de la variable auxiliaire 0 existante).

1.2. Explication mathématique

- La méthode des polygones d'influence de Thiessen

$$\text{EG(T)} = \sum_{i=1}^{n} \frac{S_1}{S} z(S_1)$$

De cette formule mathématique, je comprends qu'on fait la somme de la surface du polygone d'influence du S_1 sur la surface générale S

- La variance d'extension

$$var(\text{Z}(\sigma) - \hat{Z}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i=1}^{n} C\ (s_i - s_j) - 2\ \sum_{i=1}^{n} \frac{1}{n\,|\sigma|} \int_{\sigma}^{\cdot} C\ (s_i - s) ds + \frac{1}{|\sigma|^2} \int_{\sigma}^{\cdot} \int_{\sigma}^{\cdot} C\ (s - t) ds dt$$

C'est la variance d'estimation de σ par les échantillons $Z(s_i)$

De cette formule mathématique, je comprends qu'on fait la différence entre la variance d'estimation $Z(\sigma)$ avec la variance $\hat{Z}$ qui est la moyenne sur un domaine.

Je préfère souligner que j'ai un peu de mal à comprendre les formules mathématiques de ce papier. C'est pourquoi il m'est difficile d'en donner une explication claire.

1.3. Evaluation du papier 1

Sur le fond : pour une personne comme moi qui n'ai pas fait d'études en mathématiques ou statistiques/probabilités, les explications ont été très claires et j'ai réussi à comprendre dans les grandes lignes ce qu'ont voulu exprimer les auteurs.

Sur la forme : le papier est agréable à lire, bien structuré, aéré. Les sources ont été indiquées, mais il aurait peut-être été pertinent de mieux citer certaines sources, notamment celles de *internet* ou encore *Wikipédia* en donnant les liens url. De plus, attention aussi à quelques fautes de frappe et de fautes d'orthographe.

Dans l'ensemble, j'ai apprécié lire ce document car j'ai appris des choses que je ne connaissais pas, même si je ne vois encore comment je pourrais le mettre en pratique. Si je devais faire une petite critique, une conclusion à la fin aurait été la bienvenue.

2. Papier 2

Nom du package R : **Systèmes de conduites autonome**

Auteur du travail : Akram BENSALEM

Lien du travail sur Github :

- https://github.com/AkramBensalemPSB/PSB

2.1. Synthèse du papier 2

Ce papier a pour sujet les systèmes de conduite autonome et leurs algorithmes et formules mathématiques associés. Les voitures autonomes sont équipées de GPS, de capteurs-caméras-radars intégrés qui leur permettent de détecter des paramètres pouvant avec un impact sur la conduite. Tous ces équipements reposent sur des algorithmes et concepts mathématiques permettant aux voitures d'être autonomes.

Le principal algorithme utilisé est le *EM planner*. C'est un algorithme qui prédit les futures trajectoires et manœuvres d'une voiture. Il se base sur la géométrie différentielle ou la quadrature.

L'autre algorithme utilisé est le *dynamic programming path* qui complète le principal. Il évalue le coût de déplacement de la voiture selon le paramètre du parcours suivi et des obstacles qui ont été rencontrés.

2.2. Explication mathématique

- Le coût du du smoothness

$$C_{smooth}(f) = w_1 \int \left(f'(s)\right)^2 ds + w_2 \int \left(f''(s)\right)^2 ds + w_3 \int \left(f'''(s)\right)^2 ds$$

w représente les poids. Cette formule mathématique calcule avec les intégrales la dérivé première, seconde et troisième de s qui est le trièdre de Frenet référant à un objet et représenté par un point dans un repère.

- f'(s) = changement de cap par rapport à la trajectoire initiale
- f''(s) = courbe du chemin
- f'''(s) = dérivée de la courbe

- La fonction guidance

$$C_{guidance}(f) = \int f(s) - g(s))^2 \, ds$$

Cette formule mathématique calcule la différence entre le meilleur parcours avec l'intégral de la fonction f la situation la plus optimale que peut avoir une voiture (obstacle nulle, temps de parcours minimum) et un chemin type que peut rencontrer une voiture avec des obstacles.

2.3. Evaluation du papier 2

Sur le fond : le lecteur comprend rapidement le sujet de ce papier. L'auteur va à l'essentiel de son exposé, c'est-à-dire les algorithmes et les concepts mathématiques liés aux systèmes de la conduite autonome. Les explications sont présentes et compréhensibles.

Sur la forme : le travail sur ce travail est assez clair, aéré et rapide à lire. Les sources ont bien été données, ce qui permettent de lire le papier si on le souhaite. Attention à plusieurs fautes de frappe et d'orthographe.

Dans l'ensemble, j'ai apprécié lire ce document car il est intéressant de connaitre comment les voitures autonomes ont été réfléchies et construites avec les algorithmes. Si je devais faire une petite critique, ca serait de connaitre l'opinion de l'auteur sur le papier : qu'a-t-il pensé de ce papier, en quoi ce papier est pertinent etc.

3. Papier 3

Nom du package R : **Algorithme Génétique**

Auteur du travail : Ramya WURAOLA

Lien du travail sur Github :

- https ://github.com/RamyaHTDJ/Psb_Ramya

3.1. Synthèse du papier 3

Ce papier a pour sujet les algorithmiques génétiques et qui souhaite expliquer la création de nouvelles populations potentielles en utilisant les opérateurs évolutionnaires tels que la sélection, le croissement et la mutation.

Les fondamentaux de l'algorithmique génétique reposent sur la population, les chromosomes, les gènes, l'allèle, le génotype, le phénotype, le décodage et l'encodage, la fonction de performance ou fitness, les opérateurs génétiques. Les algorithmes génétiques s'appuient sur la théorie d'évolution des espèces de Darwin et qui reposent sur le principe de variation, d'adaptation et d'hérédité.

Pour bien construire un algorithme génétique, il est important de bien définir sa représentation qui peut être binaire, en nombres entiers ou en permutation. Puis il est nécessaire d'initialiser une population. Cette initialisation est soit aléatoire soit heuristique. Ensuite on y construit la fonction de performance qui est une fonction indiquant quelle est la bonne solution par rapport à la problématique de départ. On attribue par la suite la sélection des parents à l'algorithme génétique. Après toutes ces opérations, on effectue de croisements biologiques et des mutations. Enfin, il est nécessaire de mettre une condition qui détermine la fin de l'algorithme génétique.

3.2. Explication mathématique

A proprement parlé, il n'y a pas de formules mathématiques ou un exemple d'algorithme génétique dans ce papier. Celui-ci a surtout voulu expliquer la construction d'un algorithme génétique pouvant être exploité dans différents secteurs tels que les réseaux des neurones, le traitement d'images, ou encore résoudre des divers problèmes de gestion sur la planification.

3.3. Evaluation papier 3

Sur le fond : pour une personne comme moi qui a très peu de connaissances en génétique, les explications données sur la construction d'un algorithme génétique ont été très claires. L'auteur de ce travail a par ailleurs fait un travail d'auto-critique en finissant les avantages et les limites des algorithmes génétiques.

Sur la forme : le papier est très agréable à lire, bien structuré, aéré. Les sources ont été indiquées.

Dans l'ensemble, j'ai beaucoup apprécié lire ce document car j'ai appris des choses que je ne connaissais pas. Si je devais faire une petite critique, j'aurais aimé avoir un exemple concret d'algorithmes génétiques utilisés dans les secteurs mentionnés à la page 14, dans la section « *Champs d'application* ».

4. Papier 4

Nom du package R : **Apprentissage par Arbres de Décisions**

Auteur du travail : Antoine SERREAU, Corentin BRETONNIERE , Benjamin GUIGON

Lien du travail sur Github :

- https ://github.com/aserreau/PSB1/tree/main/Travaux%20Math%C3%A9matiques
- https ://github.com/CorentinBretonniere/CBRETONNIERE-PSBX
- https ://github.com/benjaminguigon/PSBX

4.1. Synthèse papier 4

Ce papier a pour sujet l'apprentissage par les arbres de décision. Il y en de deux catégories : (i) les arbres de régression et les arbres de classifications. Le premier prédit un résultat quantitatif et le deuxième prédit un résultat qualitatif.

(i)- l'arbre de régression est constitué de nœuds exprimé en pureté qui est calculé avec l'indice de Gini. Plus celui-ci est proche de 0, plus le nœud est pur. Il est par ailleurs possible d'évaluer la qualité de la variable dans l'arbre de décision avec la notion de coût du nœud.

(ii)- l'arbre de classifications prend aussi en compte la notion de pureté avec l'indice de Gini et du coût du nœud.

4.2. Explication mathématique

- L'ensemble d'apprentissage

(x,Y) = ($x_1, x_2, x_3, \dots\dots, x_k$, Y) avec x intputs permettant de prédit l'output Y

- L'indice de Gini

  G_i = 1- $\sum_{k=1}^{n} p^2_{i,k}$

Il calcule la probabilité de la pureté d'un nœud en soustrayant avec la probabilité d'avoir un individu de classe k dans le ième nœud.

A proprement parlé, il a peu de formules mathématiques ou un exemple d'algorithmique dans ce papier. Le but de celui-ci est de présenter rapidement la notion d'apprentissage des arbres de décisions.

4.3. Evaluation du papier 4

Sur le fond : pour une personne comme moi qui n'ai pas fait d'études en mathématiques ou statistiques/probabilités, les explications ont été très claires et j'ai réussi à comprendre dans les grandes lignes ce qu'ont voulu exprimer les auteurs.

Sur la forme : le papier est plutôt agréable à lire, structuré, aéré. Il aurait peut-être été pertinent de mieux citer les sources en mettant à la fin du rapport une bibliographie.

Dans l'ensemble, le rapport a été rapide à lire. Il explique rapidement les notions des arbres de décisions. Cependant pour un travail réalisé à 3, il aurait peut-être été intéressant de plus développer les notions en prenant un exemple concret de la vie réelle, prendre un papier de recherche qui a vraiment utilisé les arbres de décisions et appliquer dans un secteur tel que les transports ou le recherche.

5. Papier 5

Nom du package R : **Time domaine approache**

Auteur du travail : Siva CHANEMOUGAM

Lien du travail sur Github :

- https ://github.com/Siva-chane/PSBX/tree/main/math

5.1. Synthèse du papier 5

Ce papier a pour sujet le papier de H. Garnier, P.C. Young sur « *Time-domain approaches to continuous-time model identication of dynamical systems from sampled data* ». Celui-ci aborde les thématiques des fonctions continues et discrètes dans un système dynamique.

Ce papier montre qu'il est possible un modèle continu à partir d'un modèle discret en utilisant plusieurs approches : (i) la première est une approche indirecte (un ensemble de jeux de données discrets est transformé en jeux de données continus) et (ii) la deuxième est une approche directe (identification un ensemble de données continues à partir d'un modèle discret).

Pour résoudre des équations, il y a plusieurs méthodes. Il y a la méthode SVF (state variable filter -) et la méthode stochastique.

5.2. Explication mathématique

- Un mode continu à un mode discret

$$Z^N = u(tk)$$

$y(tk)_{k=1}^{N}$ = x(tk) + v(tk) équation sur la notion de bruit

Cette notion permet de passer d'un mode continu à un mode discret. Il m'est difficile de vous expliquer plus car je n'ai pas bien compris comment on est passé d'un état continu à un état discret…

- Formule de IV méthode

$\hat{\theta}_N^{IV} = \left[\sum_{k=1}^{N} \phi_f(tk)\, y_f^n(tk)\right]^{-1} \sum_{k=1}^{N} \phi_f(tk)\, \phi_f^T(tk)$

$\phi_f^T(tk) => ensemble\ des\ élements\ en\ entrée\ et\ sortie$

Il est assez difficile de comprendre la formule mathématique pour moi. Je comprendre que cette formule sert pour le θ sans le bruit. Le θ correspond aux coefficients A et B d'un système indiquant le lien entre la sortie et l'entrée

5.3. Evaluation du papier 5

Sur le fond : pour une personne comme moi qui n'ai pas fait d'études en mathématiques ou statistiques/probabilités, les explications sur les formules et les notions mathématiques ont été assez difficile. En effet, j'ai réussi à comprendre dans les grandes lignes mais superficiellement ce qu'a voulu exprimer l'auteur. Il aurait été souhaitable que l'auteur sélectionne que quelques formules mathématiques mais en donnant plus d'explications et non que de la description.

Sur la forme : le papier a été difficile lire malgré une structure dans le travail. Pour une meilleure lisibilité dans les formules mathématiques, il aurait de mieux de faire des sautes de ligne pour séparer les formules des explications. Par ailleurs, il aurait peut-être été pertinent d'avoir le lien de papier dans une bibliographie pour aider le lecteur à avoir accès plus rapidement à la source de ce travail.

Dans l'ensemble, ce papier de recherche est difficile à comprendre pour un profil comme le mien. Les explications sont plutôt abstraites et théoriques mais cela est dû au travail demandé de départ qui était d'explication des papiers de recherche avec des notions mathématiques ou algorithmiques. Si je devais faire une petite critique, une conclusion sur le papier de recherche à la fin aurait été la bienvenue et plus d'explications.

6. Evaluation de mon travail de 3 papiers de recherche par rapport aux 5 travaux sélectionnés

Je pense que sur la forme, le travail que nous avons réalisé avec mon groupe sur 3 papiers de recherche avec des concepts mathématiques est de la même qualité que les 5 travaux sélectionnés ci-dessus. Nous avons essayé de synthétiser au moins les papiers en faisant un rapport fluide et structuré à lire pour le lecteur. Nous avons aussi cité toutes nos sources et essayer d'expliquer au moins chaque notion pour que cela soit compréhensif par tous.

Sur le fond, je pense que nous avons résumé au mieux les idées des 3 papiers de recherche en faisant une synthèse des idées, en les critiquant et en les comparant entre eux.

Comme pour certains travaux sélectionnés, les concepts mathématiques peuvent vous avoir semblé léger. Cependant notre groupe en a eu pleinement conscience puisque nous avons fait un travail d'autocritique de notre rapport. Cette faiblesse mathématique s'explique notamment de nos parcours éducatifs.