Rapport Tâche 4 : détection de points aberrants

Interpolaspline

Avril 2020

1 Introduction

Il y a plusieurs méthodes différentes pour interpoler des données malgré des points aberrants. La méthode la plus intuitive est de détecter ces points et de les supprimer avant de tracer la spline. C'est cette détection de points aberrants qui a été étudiée durant la tâche 4. Trois membres du projet ont travaillé sur cette tâche : Béryl, Mohamed et Zakaria. Les méthodes ont été réparties entre ces membres du projet. Zakaria a également étudié le moyen de rendre l'étude locale (ce problème sera détaillé plus tard).

2 Objectif

L'objectif de la tâche était de rechercher et d'implémenter quelques méthodes de détection de points aberrants. Peu de temps était prévu pour cette tâche car on pensait que les méthodes seraient plutôt simples. C'est en effet le cas, néanmoins un problème s'est posé alors qu'on n'y avait pas pensé durant la planification des tâches : l'étude doit être faite localement, pour un groupe de points proches les uns des autres. En effet, si par exemple on prend la fonction identité discrétisée sur [0,10] avec un point aberrant (1,10), ce point ne sera pas detecté si l'on considère tous les points en même temps, car si celui-ci est décrété aberrant (car il se situe trop loin de la moyenne par exemple), alors (10,10) sera aussi considéré comme aberrant bien que ce ne soit pas le cas. Il faut donc trouver un moyen de séparer les points en groupes de points pas trop éloignés (exceptés éventuellement les aberrants) afin de les étudier groupe par groupe.

3 Implémentation

Nous allons détailler dans cette section les algorithmes des méthodes, puis ceux étudiés afin de créer des intervalles d'étude. Ensuite, nous allons effectuer des tests.

3.1 Algorithmes des méthodes

Les méthodes ont pour certaines des paramètres (coeff, alpha, ...). Ceux-ci sont à adapter manuellement à chaque exemple. Les estimer automatiquement paraît très compliqué. On essaira cependant de faire ça plus tard.

3.1.1 Méthode inter-quartiles

Dans cette partie, il est question de la détection des points aberrants en utilisant le 1er et 3e quartile Le premier algorithme créé prend en argument une liste ordonnée d'éléments et retourne le 1er et 3e quartile de x. Dansunsecondtemps,undeuxièmealgorithmeaétécréé,quiprendenarguments

l'intervalle et indique s'il est aberrant ou non : Un point est dit aberrant lorsqu'il est inférieur à Q1 - coeff*(Q3 - Q1) ou supérieur Q3 +coeff*(Q3 - Q1) Avec coeff un paramètre de la méthode.

3.1.2 Test de Chauvenet

Cettefonctionprendenentréeunelistesuivantunedistributiondelanormale,pourchaque élément on applique le théorème central limite et on retourne la liste de valeurs aberrantes Le test s'effectue sur un nombre qui est égale au produit de la longueur de liste par la probabilité que X soit plus grande que l'élément de liste centrée réduit, une valeur est aberrante si ce nombre est inférieur à 0.5

3.1.3 Test Tau de Thompson

ZAKARIA

3.1.4 Test de Grubbs

Cette méthode a été inventée par Frank E. Grubbs en 1969.
Pour ce test, on n'étudie que la valeur extrême (celle dont la valeur absolue de l'écart à la moyenne est la plus grande). Si plusieurs existent, on peut itérer ce test plusieurs fois tant qu'on trouve des points aberrants, en retirant le point aberrant trouvé. On compare ensuite $\frac{v_e xtreme - moyenne}{ecart-type}$ avec le seuil critique donné par le test de Grubbs : $G_{crit} = \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{\frac{\alpha}{n},n-2}}{n-2+t^2_{\frac{\alpha}{n},n-2}}}$. ($n$ est le nombre de données,$t_{a,b}$ est le résultat de la fonction quantile de Student avec un seuil de confiance a et b degrés de liberté, et α une certaine mesure de l'erreur que l'on accepte)
Si la première valeur est plus grande que le seuil, alors la méthode de Grubbs considère que le point extrême est aberrant. α est un paramètre que nous décidons de passer à la fonction. Plus celui-ci est faible, plus la chance que les points détectés comme aberrants le soient réellements (mais dans ce cas, peu sont détectés, certains points pourtant aberrants peuvent ne pas être détectés)

3.1.5 Méthode généralisée de la déviation extrême de Student

Cette méthode a été inventée par Bernard Rosner en 1983.
Ce test est une généralisation du test de Grubbs. En anglais, ce test appelé "extreme Studentized deviate" est abbrégé ESD. L'algorithme suit les étapes suivantes :
- Récupération des valeurs extrêmes (même définition que dans le paragraphe précédent)
- Comparaison des valeurs extrêmes normalisées avec le seuil critique, qui dépend du nombre de valeurs extrêmes déjà enlevées. Valeur normalisée : $\frac{v_e xtreme - moyenne}{ecart-type}$. Seuil critique, avec i le nombre de valeurs déjà enlevées : $\frac{(n-i-1)*t_{\frac{1-\alpha}{2(n-i)},n-i-2}}{\sqrt{(n-i)*(n-i-2+t^2_{\frac{1-\alpha}{2(n-i)},n-i-2})}}$ - On ne compare pas les valeurs suivantes si un point est décrété comme non aberrant : en effet, on traite les valeurs dans l'ordre de leur "extrêmitude".

3.1.6 Méthode des k plus proches voisins

Calculer pour chaque observation la distance au K plus proche voisin k-distance ; Ordonner les observations selon ces distances k-distance ; Les données aberrantes ont les plus grandes distances k-distance ; Les observations qui ont les n pourcent plus grandes distances k-distance sont des données aberrante, n étant un paramètre à fixer. Dans un premier temps, nous avons crée un algorithme qui prend en entrée une liste et un indice, et retourne une liste contenant

FIGURE 1 – Exemple de traitements des données

les distances de l'élément à la position i à aux éléments de la liste. Ensuite, un deuxième qui comme la précédente prend en entrée une liste, un indice et un entier k et retourne la k-distance de l'élément à la position indice qui représente la moyenne ses k petites distances. Enfin,undernierquielleprenduneliste,unentierketunentierncommeindiquéaudessus, retourne la liste contenant les valeurs de la liste qui ont les n pourcent plus grande k-distance.

3.2 Algorithmes de création d'intervalles

3.3 tests

Les tests ont été réalisés sur la définition des points aberrant suivante :
Un point est dit aberrant lorsqu'il n'appartient pas à l'intervalle

$$[Q_1 - 1.5 * (Q_3 - Q_1), Q_3 + 1.5 * (Q_3 - Q_1)]$$

Le résultat est visible dans la figure 1.

3.4 Sources

`https://fr.wikipedia.org/wiki/Donn%C3%A9e_aberrante` : beaucoup de méthodes
`https://lemakistatheux.wordpress.com/category/tests-statistique-indices-de-liaison-et-c`
`les-tests-pour-la-detection-doutliers/` Plusieurs méthodes expliquées permettant de détecter les outliers (Tukey, Test Q de Dixon, Chauvenet, Grubbs, Tietjen-Moore, Déviation extrême généralisée Student, modifié de Thompson, critère de Peirce) + des exemples
`https://en.wikipedia.org/wiki/Grubbs%27s_test_for_outliers` test de Grubbs
`https://ellistat.com/guide-dutilisateur/statistiques-descriptives/tests-de-valeurs-aber`
`test-de-grubb/` test de Grubbs (pratique seulement, pas de théorie)