Evaluation des dossiers sur les mathématiques du Big Data

Les « 12 travaux d'AstéRix »

Daif Hakim

31/01/2021

Arbres de décisions

1. Critères d'évaluation

Dans ce dossier, nous allons parcourir et faire une évaluation des travaux de mes camarades en mathématiques du Big Data:

Nos critères seront :

1. Visuel et l'organisation du dossier: Es ce que la mise en page est propre, agréable à lire ou non ? Le travail est-il organisé (chapitres, sections, sous-sections, Parties)

2. Qualité du latex et des formules mathématiques : Es ce que les formules sont claires ou pas ? Es ce que le dossier contient le code Latex ou le rapport sous format Rmd ? **3. Compréhension de l'idée générale:** Les Auteurs arrivent-ils à transmettre l'idée générale ?

4. Explication et compréhension des formules mathématique et des concepts fondamentaux: Es ce que les formules et les notions sont bien expliquées ?

5. Difficulté et originalité du thème : Le sujet abordé est-il original/difficile ou pas ?

2. Lien vers le document commenté

l'un des travaux que nous allons analyser dans ce dossier est celui de Antoine SERREAU, Benjamin GUIGON et Corentin BRETONNIERE sur le sujet des arbres de décisions.

Les liens vers leurs Github: **https://github.com/CorentinBretonniere/CBRETONNIERE-PSBX**

Le dossier en question est **Article_Arbre_décision.pdf**

3. Synthèse de la présentation

Les arbres de décisions modélisent une hiérarchie de test pour prendre une décision ou prédire un résultat en fonction des expériences précédentes.

Les auteurs reviennent sur l'aspect mathématique des deux types d'arbres : régression et classification, notamment la notion de pureté et coût du nœud.

Ils commencent tout d'abord par expliquer l'arbre de régression en s'appuyant sur son aspect mathématique notamment sur la notion de pureté et le cout du noeud
Ils illustrent leur explication par un exemple inspiré du travail de Christophe Chesnau intitulé "Introduction aux arbres de décisions", le dataset utilisé dans cet exemple est : Iris.

Ensuite ils abordent le deuxième type d'arbres de décision qui la classification, ils ont procédé de la même façon que le précédent type mais d'une façon un peu moins détaillée, vu qu'ils ont eu à faire aux mêmes notions notamment, l'Indice de Gini et cout du nœud

Ils finissent leur travail en indiquant les limites des arbres de décision : -Problème du NP complet. -Instabilité . -Problème de sur-apprentissage ; Les auteurs précisent aussi que Random Forest peut être une solution aux problèmes d'instabilités et de surapprentissage, .

4. Explication des formules

Afin d'expliquer les arbres de décision, les auteurs se sont focalisés sur l'explication des deux notions propres au nœud qui sont la pureté d'un nœud mesurée par l'indice de génie et la notion de cout du nœud qui mesure à quel point le choix de la variable de décision est bon

La première formule permet de calculer la pureté d'un noeud, un noeud est dit pur si tous les individus associés sont de la même classe et que la valeur est 0. La formule est donc basée sur la probabilité d'avoir un individu d'une classe k parmi la population au noeud i.

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{1}$$

La seconde formule concerne le coût du noeud, celle-ci inclut la formule de la pureté d'un noeud ce qui implique que plus les noeuds sont purs plus le coût est faible.

$$J(k) = (\frac{m_{gauche}}{m})G_{gauche} + (\frac{m_{droite}}{m})G_{droite} \tag{2}$$

5. Evaluation

1. Visuel et l'organisation du dossier: Le visuel est bon, travail très organisé (section, titres, exemples) cependant on ne trouve pas les chunks de code R, ou bien un fichier latex

2. Qualité du latex et des formules mathématiques : les formules sont très claires et concises cependant on ne trouve pas dans ce dossier le code latex ce qui nous permet pas d'évaluer leur maitrise du code latex.

3. Compréhension de l'idée générale : les auteurs métrisent bien les différentes notions, le travail est très clair et concis, bien organisé, toutes les parties sont bien expliquées cela facilite énormément la compréhension de l'idée générale

4. Explication et compréhension des formules mathématique et des concepts fondamentaux: Des formules très claire, facile à comprendre, cependant, on constate le manque des détails, rien qu'à travers les formules et l'enchainement des idées on peut déduire que les notions fondamentales sont bien comprises ce qui est très bien

5. Difficulté et originalité du thème : le sujet abordé est ordinaire il a été traité plusieurs fois dans le cours , en termes de difficultés, les notions introduites sont basiques, cependant ils auraient pu développer encore plus les mieux mathématique et détailler encore plus la partie Arbres de Classification

6. Conclusion

En conclusion, on trouve que le rapport est agréable à la lecture car très organisé et les notions fondamentales sont bien comprises par les auteurs . cependant, ils auraient pu faire mieux en essayant d'expliquer un peu plus les différentes parties et notions, et pourquoi pas introduire plus d'exemples. Autre point à améliorer et qui est très important : pas de chunks ou code R