Decison Tree

Zakaria RIDADARAJAT

02/02/2021

Le travail que nous allons analyser dans ce dossier est celui de Corentin BRETONNIERE sur le sujet des arbres de décisions.

Les liens vers leurs Github:

https://github.com/CorentinBretonniere/CBRETONNIERE-PSBX

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.

L'auteur revient sur l'aspect mathématique des deux types d'arbres : régression et classification, notamment la notion de coût du nœud ET pureté.

Tout d'abord il explique ce qu"arbre de régression, en s'appuyant sur son aspect mathématique notamment sur la notion de pureté et le cout du nœud. Il illustre aussi 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 il expose le deuxième type d'arbres de décision qui la classification, il a procédé de la même manière que l'arbre de régression. Néanmoins cette partie a été moins détaillé.

Il termine son rapport par mentionner quelques limites des arbres de décision tels que : Instabilité et le Problème de sur-apprentissage ;

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 nœud, un nœud 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.

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

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

$$J(k) = \left(\frac{m_{\text{gauche}}}{m}\right) G_{\text{gauche}} + \left(\frac{m_{\text{droite}}}{m}\right) G_{\text{droite}}$$

1) lisibilité du rapport :

Le visuel est bon, travail très organisé (section, titres, exemples)

5) Qualité du LaTeX

Les formules sont très claires et concises cependant on ne trouve pas dans ce dossier le code latex ce qui ne nous permet pas d'évaluer leur maitrise du code latex.

2) qualité du rapport :

Le travail est très clair et concis, bien organisé, toutes les parties sont bien expliquées.

3) Aspect didactique

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

4) Bibliographie :

Effectivement, les auteurs ont mentionné la bibliographie à la fin du rapport pour les gens qui veulent approfondir un peu plus sur le sujet.

6. Conclusion

Personnellement, je trouve que le rapport est agréable à la lecture car très organisé et les notions fondamentales sont bien comprises par l'auteur.

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 et aussi exposer les chunks ou le code R.