TME 1 - Arbres de décision, sélection de modèles

CHERCHOUR Lièce & DIEZ Marie

Avec l'aide de THAUVIN Dao & STERKERS Luc

March 21, 2021

1 Entropie

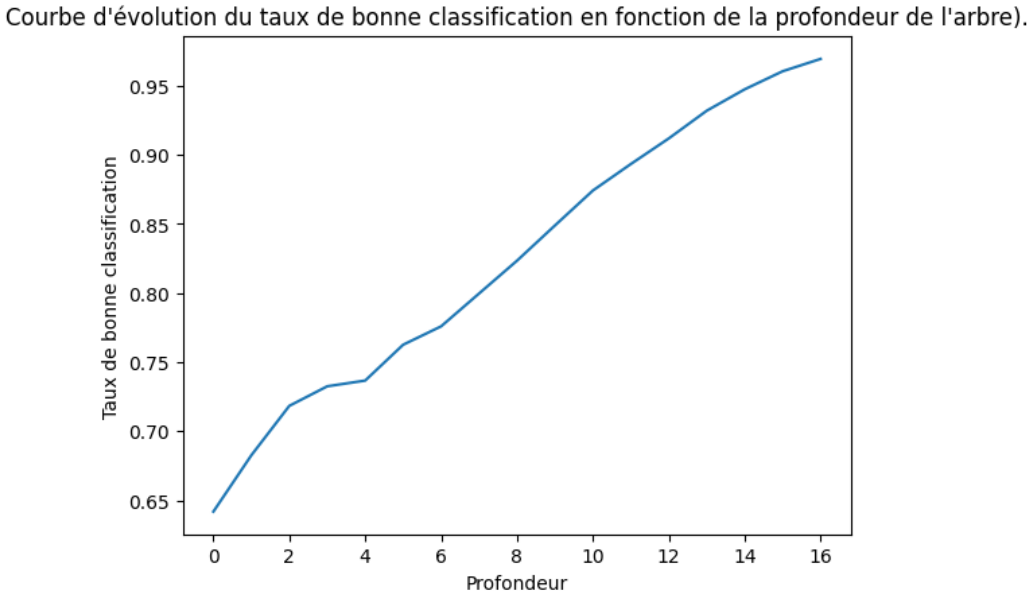
L'entropie permet de définir le désordre d'un système, l'entropie conditionnelle nous permet ici de calculer l'entropie conditionnellement à un attribut, de cette manière on peux trouver quel attribut minimise l'entropie, c'est-à-dire l'attribut qui discrimine au mieux les classes. Cela nous permettra de choisir à chaque étape, sur quel attribut faire le test de décision. Une entropie égale à 0 signifie qu'il n'y a aucun désordre sur les données par exemple tous les labels sont égaux {1,1,1,1,1,1}, à l'inverse si l'entropie vaut 1 cela signifie que le désordre est maximale {0,0,0,1,1,1} par exemple en classification binaire. On peut aussi calculer le gain d'information :

$$I = H(Y) - H(Y|P) = -\sum_y pylog(py) + \sum_y p(y|Pi)log(p(y|Pi))$$

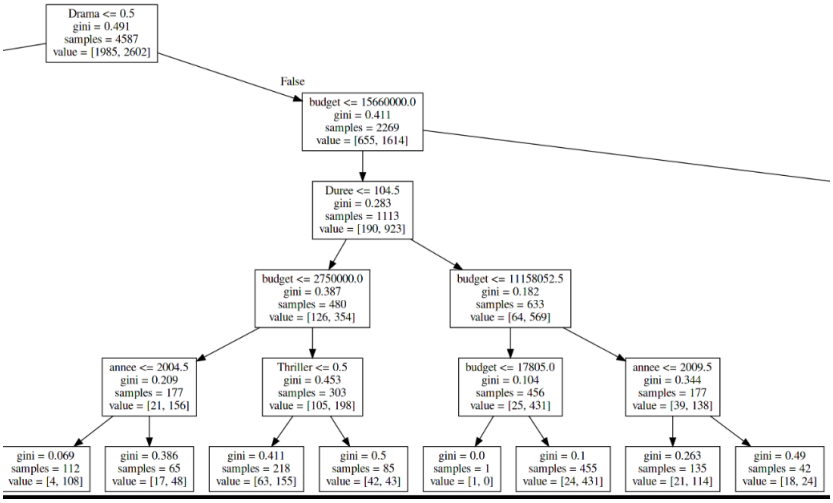
De cette manière nous obtenu comme meilleur attribut pour la première partition : "Drama" id=17 avec pour gain d'information : 0.06

2 Arbres de décision

Le graphe si-dessous permet de visualiser le taux de bonne classification en fonction de la profondeur de notre arbre :



En fonction de la profondeur que l'on souhaite obtenir dans notre arbre, le nombre d'éléments séparés à chaque niveau varie.



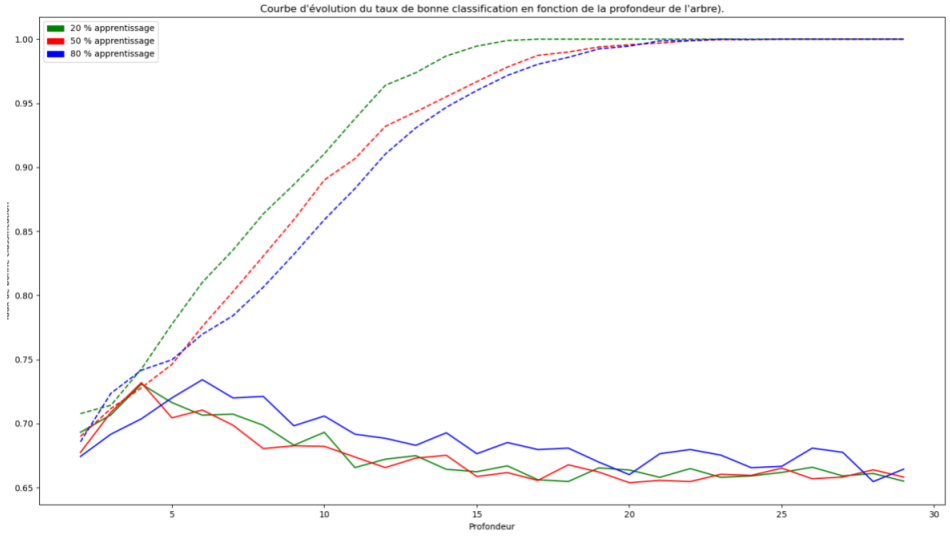
On observe sur cette partie d'un arbre de décision, que plus l'on descend dans l'arbre, moins un noeud contient d'exemple. Cela est normale car on rajoute des tests et on tend donc à séparé de plus en plus nos exemples.

Dans le graphe plus haut, on peut facilement observé une différence entre le taux de bonne classification selon la profondeur données. Notre score de classification tend vers des valeurs de plus en plus proche de 1, c'est tout à fait normal car notre arbre va séparer de moins en moins de données jusqu'à obtenir des noeuds avec très peu voir un seul exemple.

Le problème est que dans cette exécution, nous n'avons pas séparé données de test et données d'apprentissage, tous nos tests ont était effectué sur seulement les données d'apprentissages. Nous n'avons donc pas un indicateur fiable des performances de notre algorithme. Une façon d'obtenir un indicateur plus fiable serait de séparé nos données en des données d'apprentissage et de tests.

3 Sur et sous apprentissage

Il est important de faire attentions aux problèmes de sous apprentissage et de sur apprentissage pour obtenir de bon résultats sur nos données de test.

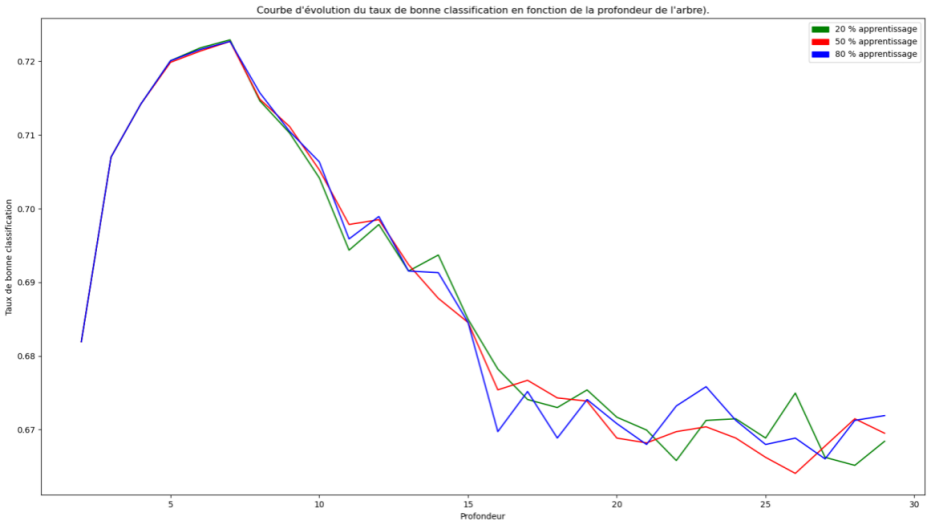


Les tracés en pointillés représente le taux de bonne classification sur les données d'apprentissages et les courbes pleines sur les données de test.

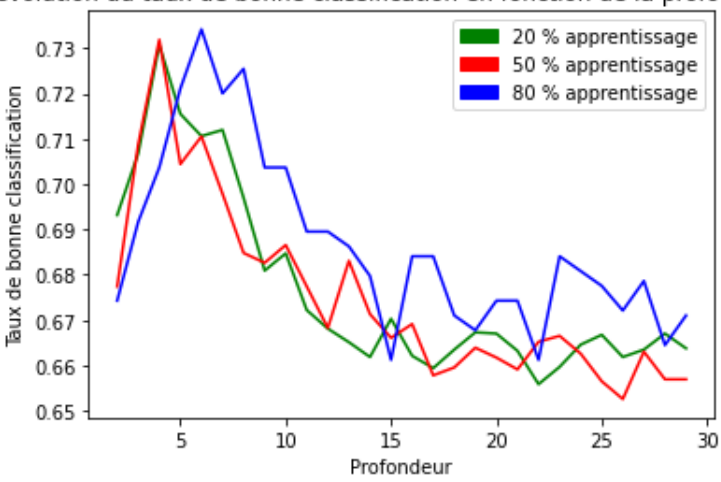
Plus la profondeur augmente plus les erreurs de classification sont faibles sur les données d'apprentissage, en effet nous avons appris sur ces données donc le modèle est parfaitement adapté à celle-ci, cependant à

partir d'une certaine profondeur on remarque une diminution du taux de bonne classification en test, on arrive à un point où le modèle se spécialise sur les données d'apprentissage et n'est donc pas robuste face à de nouvelles données, c'est du sur-apprentissage. Dans le cas inverse si la profondeur est trop faible, on observe le phénomène de sous-apprentissage, le modèle n'est pas assez performant pour la classification de données, que ce soit en test ou en apprentissage. Il faut alors trouver la profondeur optimale en prenant compte de ces phénomènes.

Nos résultats peuvent ne pas être tout à fait fiable si les données ne sont pas assez nombreuses, dans quel cas pour avoir meilleurs performances ont peut mettre en place un système de validation croisée.



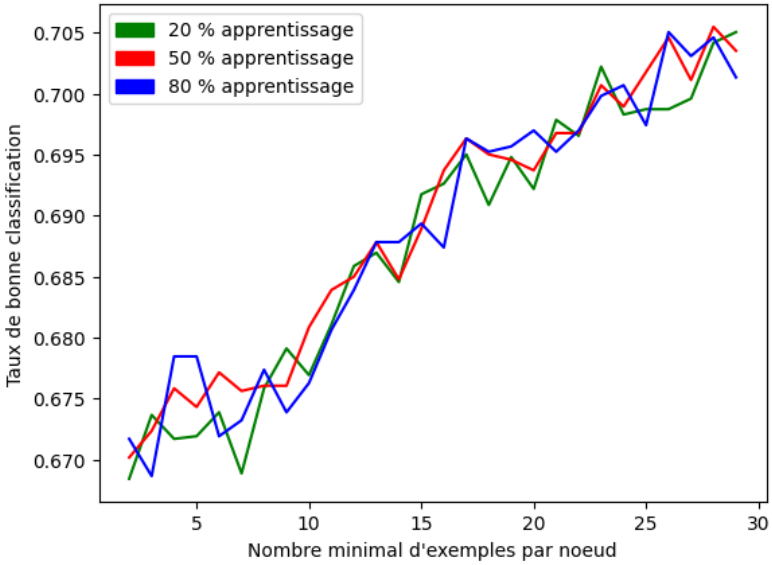
Courbe d'évolution du taux de bonne classification en fonction de la profondeur de l'arbre).



La première image corresponds aux taux de bonne classification sur les données de test avec validation croisée et la seconde sans validation croisée, on peux remarquer qu'il n'y a qu'une différence insignifiante entre les deux, nous avons sans doutes suffisamment de données pour avoir de bonne performances sans avoir besoin de mettre en place de validation croisée.

Nous avons également effectué des tests avec d'autres hyper-paramètres (tel que le gain d'entropie minimal et le nombre minimal d'exemples par noeud), voici les graphes associés :

Courbe d'évolution du taux de bonne classification en fonction du gain d'entropie minimal).



Courbe d'évolution du taux de bonne classification en fonction du nombre min d'exemple par noeud).