APPENDIX

BELKHITER Yannis, FORAY Léo-Paul, MU Maxime, TEXIER Lucas

5 avril 2023

Résumé

L'Apprentissage Automatique est un vaste domaine regroupant plusieurs théories, ayant chacune leurs modèles respectifs. Les Arbres de décisions et les Random Forest sont des modèles faisant partie de ce domaine du Machine Learning, s'adaptant autant à des problématiques de régression que de classification.

Table des matières

1 Arbre de décision

1.1 Introduction

Les arbres de décision sont une méthode d'apprentissage supervisée utilisée pour la modélisation prédictive dans un contexte de régression ou de classification. Les arbres de décision sont un type de modèle qui utilise un graphique arborescent ou un modèle de décisions. Il s'agit de l'un des algorithmes d'apprentissage automatique les plus populaires et les plus efficaces. Dans ce rapport, nous allons examiner les différents aspects des arbres de décision, de leur construction à leur utilisation en passant par leur évaluation. Nous allons également discuter de leurs avantages et de leurs inconvénients, ainsi que de leurs applications dans différents domaines. Enfin, nous allons fournir des exemples pratiques pour aider le lecteur à mieux comprendre leur fonctionnement.

1.2 Les arbres de décision : définition

Comme présenté en introduction, un arbre de décision est un modèle composé de nœuds et d'arêtes. Il s'agit finalement d'un graphe orienté (d'un arbre) qui se compose de nœuds, dont un nœud racine sans arêtes entrantes, ainsi que de nœuds internes de test et de nœuds feuilles (également connus sous le nom de nœuds terminaux ou décisionnels). Les nœuds représentent les décisions ou les événements, tandis que les arêtes représentent les conséquences ou les résultats de ces décisions.

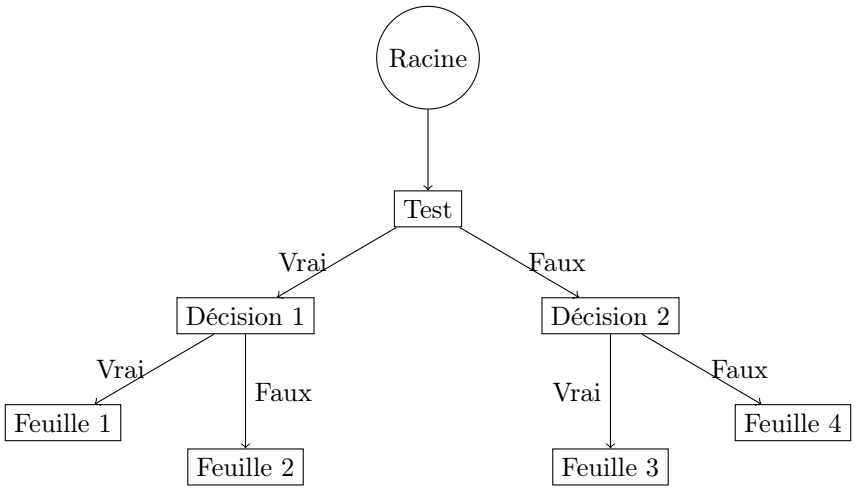

FIGURE 1 – Exemple d'un arbre de décision

1.3 Construction d'un arbre de décision

Pour construire un arbre de décision, on commence par sélectionner l'attribut qui divise le mieux l'ensemble des données en fonction des classes ou des valeurs à prédire. Cette sélection est effectuée en utilisant une mesure d'impureté qui évalue la répartition des classes ou des valeurs de la variable à prédire pour chaque partition. Les mesures d'impureté les plus courantes sont l'entropie et le coefficient de Gini.

1.3.1 Tests d'impuretés

La décision des arbres se base sur les nœuds de décision. Cette dernière est basée sur l'attribut qui divise le mieux l'ensemble comme nous l'avons expliqué. Mais comment choisit-on l'attribut séparateur ? Par un test d'impureté.

Une mesure d'impureté est une mesure qui permet d'évaluer l'homogénéité d'un ensemble d'échantillons. Dans le contexte des arbres de décision, ces mesures sont utilisées pour évaluer la qualité de chaque division potentielle du jeu de données en sous-ensembles. L'objectif est de trouver la division qui réduit le plus l'impureté des sous-ensembles résultants. En effet, en utilisant des mesures d'impureté appropriées, les arbres de décision peuvent être construits de manière à minimiser l'erreur de

classification ou l'erreur de prédiction, en divisant les données de manière à ce que les sous-ensembles résultants soient les plus homogènes possibles.

En résumé, plus les données sont divisées de façon homogènes dans un arbre de décision, et meilleurs seront les résultats de sorties

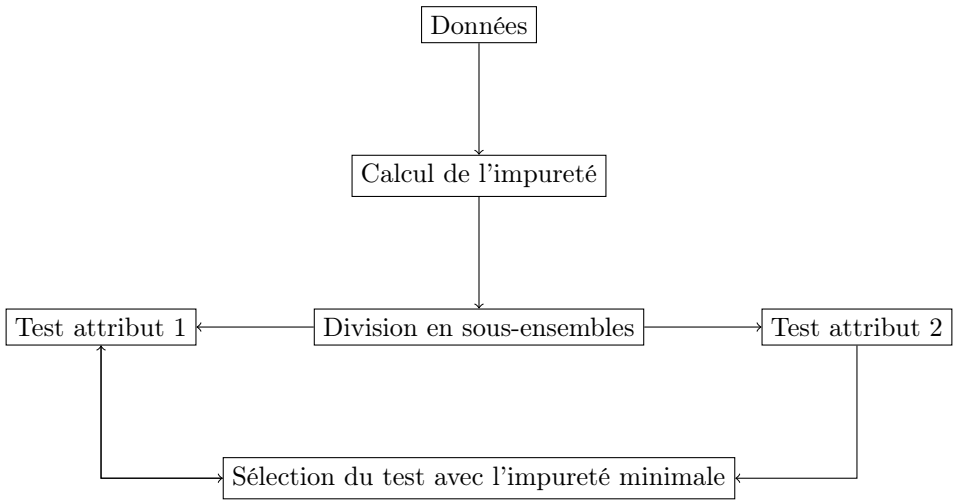

FIGURE 2 – Processus de sélection de l'attribut séparateur basé sur l'impureté (par Récursivité)

PS : *Il existe plusieurs mesures d'impureté couramment utilisées pour les arbres de décision, telles que l'indice de Gini et l'entropie. L'indice de Gini mesure la probabilité que deux éléments choisis au hasard dans un sous-ensemble soient mal classés s'ils sont étiquetés au hasard selon la distribution des classes dans ce sous-ensemble. L'entropie, quant à elle, mesure la quantité d'incertitude associée à la distribution des classes dans un sous-ensemble.*

1.3.2 Récursivité et structure du modèle

Une fois l'attribut sélectionné, on divise l'ensemble des données en fonction des valeurs de cet attribut, créant ainsi des sous-ensembles. On répète ensuite ce processus de manière récursive pour chaque sous-ensemble jusqu'à ce que chaque sous-ensemble ne puisse plus être divisé en fonction d'un critère d'arrêt prédéfini, tel que le nombre minimal d'observations dans un nœud ou la profondeur maximale de l'arbre.

En somme, pour chaque nouvelle observation, on suit le chemin correspondant dans l'arbre de décision en fonction des valeurs de ses attributs, jusqu'à atteindre un nœud terminal. Le nœud terminal est alors associé à une valeur de sortie prédite, qui est renvoyée comme résultat de la prédiction. Dans le cas de la régression, les nœuds terminaux sont souvent calculés comme la moyenne des valeurs de sortie des observations correspondantes. Dans le cas de la classification, les nœuds terminaux sont souvent associés à une classe majoritaire des observations correspondantes.

1.3.3 Entraînement du modèle

Pour entraîner l'arbre de décision, on utilise un ensemble de données labellisées d'entraînement , qui sont des données avec lesquelles le modèle est construit. Le modèle est ajusté aux données d'entraînement en sélectionnant les attributs et les seuils qui divisent le mieux l'ensemble des données en fonction des classes ou des valeurs à prédire. Les valeurs prédites sont comparées aux labels associées aux données du dataset d'entraînement, et un erreur est introduite à partir d'une fonction d'erreur (MSE, MAE, ou encore RMSE). C'est cette erreur qui est ensuite utilisée pour modifier les valeurs de seuils, qui sont les valeurs de décision de chaque nœud. L'objectif ensuite classique pour ce type de modèle est de minimiser la fonction d'erreur, c'est-à-dire déterminer les valeurs de seuils donnant une erreur minimale sur le dataset d'entraînement.

Nous allons différencier ce calcul d'erreur, et de prédiction selon les deux problèmes (classification et régression).

Classification : La classification vise à attribuer aux observations une classe parmi un nombre fini de classe. Pour se faire, on parcourt l'arbre de décision, et on note tous les résultats des nœuds terminaux. Le résultat de l'arbre de décision pour la donnée d'entrée sera la classe la plus occurrente de la liste des classes prédites.

Régression : Pour prédire une valeur numérique, on suit le même processus que pour la classification, mais on note les valeurs numériques obtenues pour chaque feuilles (nœuds terminants). Ensuite, on calcule la moyenne ou la médiane (cela dépend du type de données et du contexte de l'application) des valeurs obtenues, qui constitue la valeur prédite de notre modèle.

En résumé, la construction d'un arbre de décision implique la sélection des attributs et des seuils qui divisent le mieux les données en fonction des classes ou des valeurs à prédire, tandis que l'entraînement de l'arbre de décision implique l'ajustement du modèle aux données d'entraînement. Une fois l'arbre de décision entraîné, il peut être utilisé pour classer de nouvelles observations ou prédire des valeurs numériques en suivant le chemin correspondant aux résultats des tests dans l'arbre.

1.4 Evaluation des arbres de décision

Pour évaluer la qualité de l'arbre de décision, on utilise un ensemble de données de test, qui sont des données indépendantes des données d'entraînement et qui sont utilisées pour évaluer les performances du modèle en termes de précision de classification ou de prédiction. La validation classique vise à comparer les valeurs prédites avec les labels associés aux données de tests. Néanmoins, on peut aussi améliorer la qualité du processus d'évaluation en séparant notre dataset par validation croisée par exemple.

1.5 Utilisation des arbres de décision

Comme expliqué dans les parties précédentes, les nœuds de l'arbre en fonction des résultats des tests, on peut arriver à une feuille qui représente la décision finale, telle que la classification d'une observation ou la prédiction d'une valeur numérique.

Pour l'application de la régression, l'arbre de décision est utilisé pour prédire une valeur numérique continue en utilisant des variables d'entrée continues ou catégorielles. Par exemple, un arbre de décision peut être utilisé pour prédire le prix d'une maison en fonction de variables telles que la superficie, le nombre de chambres, la localisation géographique, etc.

Pour l'application de la classification, l'arbre de décision est utilisé pour classer des observations dans des catégories prédéfinies en utilisant des variables d'entrée continues ou catégorielles. Par exemple, un arbre de décision peut être utilisé pour classer des animaux sur une image (ex : chat, chien, hamster, etc...)

En résumé, l'arbre de décision est un outil puissant et flexible qui peut être utilisé pour la régression ou la classification en fonction des données d'entrée et des objectifs de la modélisation.

1.6 Avantages et inconvénients des arbres de décision

Voici quelques avantages et inconvénients des arbres de décision :

1.6.1 Avantages des arbres de décision

- Les arbres de décision sont faciles à comprendre et à interpréter, même pour les personnes qui ne sont pas spécialisées dans le domaine de l'apprentissage automatique. Ils permettent également d'identifier facilement les variables importantes qui influencent la décision.
- Les arbres de décision peuvent gérer les données manquantes en les classant dans une catégorie distincte.
- Les arbres de décision peuvent produire des résultats relativement bons par rapport à d'autres méthodes d'apprentissage automatique.

1.6.2 Inconvénients des arbres de décision

— Les arbres de décision peuvent facilement sur-ajuster les données d'entraînement, ce qui signifie qu'ils peuvent devenir trop complexes et perdre en généralité.

— Les arbres de décision sont souvent moins précis que la régression linéaire par exemple lorsque la relation entre les variables est linéaire ou presque linéaire.

— Les arbres de décision peuvent être sensibles aux données manquantes ou aux valeurs aberrantes, ce qui peut conduire à des prédictions inexactes.

1.7 Conclusion

En conclusion, les arbres de décision sont un algorithme d'apprentissage automatique largement utilisé et efficace, qui présente de nombreux cas d'utilisation dans divers secteurs. Les arbres de décision sont faciles à comprendre, à visualiser et à mettre en œuvre, ce qui en fait une option intéressante pour les parties prenantes qui ne sont pas familiarisées avec l'apprentissage automatique. Les arbres de décision ont leurs limites, comme le risque de surajustement et d'instabilité, mais ils restent un outil essentiel pour la modélisation prédictive et les tâches de classification.

1.8 Références bibliographiques

1. Cours de M. Sutton-Charani - Statistical learning - (2IA- Data science- Advanced statistics) https://campus2.mines-ales.fr/mod/resource/view.php?id=29967
2. Jijo, Bahzad & Mohsin Abdulazeez, Adnan. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. 20-28.
3. Rokach, Lior, & Maimon, Oded. (2005). Decision Trees. 10.1007/0-387-25465-X_9.
4. What is Decision tree ? - An introduction to machine learning algorithms (Aviral Bhardwaj) https://iaviral.medium.com/what-is-decision-tree-an-introduction-to-machine-learning-algorithms
5. CS 446 Machine Learning Fall 2016 SEP 8, 2016 Decision Trees Professor : Dan Roth, Scribe : Ben Zhou, C. Cervantes, University of Pennsylvania

2 Random Forest

2.1 Introduction

Les modèles Random Forest sont des variantes du modèle d'arbre de décision. Tout comme le modèle présenté précédemment, Random Forest est une méthode d'apprentissage supervisée utilisée pour la modélisation prédictive (régression) et la classification. Concrètement, c'est une combinaison de plusieurs arbres de décisions dépendants des valeurs d'un vecteur aléatoire échantillonné indépendamment et avec la même distribution pour tous les arbres de la forêt. Tout comme la présentation du modèle d'arbre de décision, nous allons dans un premier temps définir le modèle et sa construction, puis expliciter son entraînement, puis son évaluation. Pour finir, nous donnerons des exemples d'application du modèle, puis nous conclurons en essayant de ressortir les avantages et inconvénients comparé aux autres modèles classiques.

2.2 Les modèles Random Forest : définition

Comme expliqué en introduction, le modèle Random Forest est basé sur l'implémentation de plusieurs arbres de décisions. L'objectif principal du modèle est de faire ressortir la classe la plus "populaire", en générant plusieurs arbres de décisions centrées sur des caractéristiques différentes puis de les comparer aux sous-modèles générés entre eux. La génération des arbres de décision se base sur une sélection aléatoire des meilleurs paramètres, mais aussi sur des échantillons de valeurs de tailles différentes choisies aléatoirement pour chaque arbre.

2.3 Construction d'un arbre de décision

La construction d'un modèle Random Forest consiste en la création d'un nombre prédéfini d'arbres de décision. Chacun ce ces arbres sera créé à partir de deux méthodes le rendant unique. Tout d'abord, le Feature Sampling, qui permet la sélection aléatoire de caractéristiques parmi l'ensemble des caractéristiques initiales. Vient ensuite le tree bagging, c'est-à-dire la restriction à un échantillon de valeurs prises aléatoirement parmi l'ensemble de valeur initial. Chaque arbre est ensuite entraîné sur son propre échantillon, sans limitation ni uniformisation des tailles.

Les étapes de construction des arbres de descriptions sont décrites dans la partie qui est réservée dans ce modèle.

Une fois notre forêt entraînée, elle peut être utilisée pour traiter des problématiques de régression comme de classification. Pour chaque nouvelle observation, on demande à chacun de nos arbres de sélection de déterminer la valeur qui lui est associée. Ceux-ci vont identifier par descente successive la classe ou estimer la valeur associée à cette entrée. Dès lors, chaque arbre possède un poids similaire pour déterminer la classe majoritaire ou la valeur moyenne.

En résumé, la construction d'une forêt aléatoire consiste en la création de n arbres de décisions chacun entraînés sur des sous ensembles distincts de caractéristique mais aussi de valeurs. Le processus de prédiction ou choix de la classe se fait ensuite par vote ou moyenne des résultats de chacun des arbres.

2.4 Evaluation des modèles Random Forest

Pour évaluer la qualité des forêts aléatoires, on utilise un ensemble de données de test, qui sont des données indépendantes des données d'entraînement et qui sont utilisées pour évaluer les performances du modèle en termes de précision de classification ou de prédiction. La validation classique vise à comparer les valeurs prédites avec les labels associés aux données de tests. Néanmoins, on peut aussi améliorer la qualité du processus en séparant notre dataset par validation croisée par exemple.

2.5 Utilisation des arbres de décision

Comme expliqué dans les parties précédentes, une forêt aléatoire est avant tout un ensemble d'arbres de décision, et peut donc traiter des problématiques de classification ainsi que de régression.

Pour l'application de la régression, les arbres de décision sont utilisés pour prédire des valeurs numériques continues en utilisant des variables d'entrée continues ou catégorielles. On obtient alors par moyenne de ces valeurs une valeur numérique qui sera notre prédiction finale. Par exemple, une

forêt aléatoire peut être utilisée par des négociants en bourse pour prédire l'évolution de certaines côtes.

Pour l'application de la classification, les arbres de décision sont utilisés pour prédire des valeurs numériques catégorielles en utilisant des variables d'entrée continues ou catégorielles. On obtient alors par majorité la classe retenue. Par exemple, une forêt aléatoire peut être utilisée par les banques pour déterminer quels clients seront les plus propices à rembourser un crédit dans les temps.

En résumé, les forêts aléatoires sont des outils performants et polyvalents qui ont pour but de combler certaines lacunes des arbres de décisions.

2.6 Avantages et inconvénients des forêts aléatoires

Tout deux performants dans le cadre de régression ainsi que de classification, les modèles Decision tree et Random Forest ont tous deux des utilités différentes, nous allons donc essayer de les comparer pour comprendre ce qui les différencies l'un de l'autre :

2.6.1 Avantages des forêts aléatoires

— Les arbres de décisions sont dépendants du jeu de donnés et peuvent varier fortement par l'ajout/suppression de certaines instances du jeu d'entraînement (overfitting). Au contraire, les sélections aléatoires et l'utilisation de la moyenne ou de la majorité rend les forêts aléatoires moins sensibles aux variations du jeu de données.

— Les forêts aléatoires montrent de meilleurs résultats sur des jeux de données de taille importante, avec un nombre de caractéristiques semblables. Cette supériorité concerne la précision, le rappel ainsi que la proportion de bonnes réponses .

2.6.2 Inconvénients des forêts aléatoires

— Sur de petits jeux de données, les forêts aléatoires sont généralement plus lentes et n'obtiennent pas de résultats significativement meilleurs qu'un arbre de décisions

— Les arbres de décision sont plus faciles à lire et à interpréter.

2.7 Conclusion

Les forêts aléatoires sont par construction une version améliorée du modèle arbre de décision, ce qui en fait un des modèles les plus performants dans de nombreux domaines de pointe tels celui des banques et de la finance, pour lesquels on travaille sur des grands jeux de données. Ses forces sont cependant coûteuses en énergie, son intérêt est donc moindre sur des problèmes plus légers, pour lesquels un arbre de décision sera plus rapide et compréhensible.

2.8 Références bibliographiques

1. Cours de M. Sutton-Charani - Statistical learning - (2IA- Data science- Advanced statistics) https://campus2.mines-ales.fr/mod/resource/view.php?id=29967
2. Ali, Jehad Khan, Rehanullah Ahmad, Nasir Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues (IJCSI), consulted on 03/04/2023. https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees
3. Gérard Biau. 2012. Analysis of a random forests model. J. Mach. Learn. Res. 13, null (3/1/2012), 1063–1095, consulté le 03/04/2023, https://dl.acm.org/doi/10.5555/2188385.2343682
4. Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32, consulté le 03/04/2023. https://doi.org/10.1023/A:1010933404324