ARBRES DE DECISION
NOTIONS THEORIQUES

TABLE DES MATIERES

I. INTRODUCTION

Les arbres de décision sont des méthodes d'apprentissage non paramétriques utilisées pour des problèmes de classification **OU** de régression.

L'objectif est de créer un modèle qui prédit les valeurs de la variable cible, en se basant sur un ensemble de séquences de règles de décision déduites à partir des données d'apprentissage.

Ces modèles sont intéressants pour extraire de l'information d'une source de données obscure. Un arbre de décision permet d'isoler les propriétés ou features qui apportent le plus d'information pour déterminer la classe de la variable de sortie. Il s'agit donc de trouver dans une énorme quantité de données les questions qu'il est judicieux de poser afin de prédire la variable de la sortie.

Les arbres de décision ont plusieurs avantages qui rendent intéressants dans des contextes où il est utile de comprendre la séquence de décisions prise par le modèle :

- Ils sont simples à comprendre et à visualiser ;
- Ils nécessitent peu de préparation des données ;
- Le coût d'utilisation des arbres est logarithmique (pas gourmand en temps de calcul) ;
- Modèle en « boîte blanche » (le résultat est facile à conceptualiser et à visualiser)

Cependant, ces modèles présentent deux inconvénients majeurs :

- Sur-apprentissage : parfois les arbres générés sont trop complexes et généralisent mal.
  Par contre, choisir des bonnes valeurs pour les paramètres de profondeur maximale et le nombre minimal d'exemples par feuille permet d'éviter ce problème.
- Il peut arriver que les arbres générés ne soient pas équilibrés. Il est donc recommandé d'ajuster la base de données avant la construction, pour éviter qu'une classe domine largement les autres (en termes de nombre d'exemples d'apprentissage).

II. PRINCIPES

Le but général d'un arbre de décision est d'expliquer une valeur à partir d'une série de variables discrètes ou continues. On est donc dans un cas très classique de matrice X avec m observations et n variables, associée à un vecteur Y à expliquer. Les valeurs de Y peuvent être de deux sortes :

- Continues : on parle d'arbre de régression
- Qualitatives : on parle d'arbre de classification

Ces méthodes inductives présentent de nombreux atouts : elles sont assez **performantes**, **non paramétriques** et **non linéaires**.

Dans le principe, elles vont **partitionner les individus en produisant des groupes d'individus les plus homogènes possible du point de vue de la variable à prédire**, en tenant compte d'une hiérarchie de la capacité prédictive des variables considérées. Cette hiérarchie permet de visualiser les résultats dans un arbre et de constituer des **règles** explicatives explicites, orientées métier.

Il existe deux familles d'algorithmes de construction des arbres de décision :

- Divide and conquer
  - ID3 (Quinlan 1979)
  - CART (Breiman, et al., 1984)
  - ASSISTANT (Bratko 1984)
  - C4.5 (Quinlan 1986)
- Cover and diffférentiate
  - CN2 (Clark & Niblett 1983)
  - AQnn (Michalski et al., 1986)

1. DEFINITION DES REGLES

Pour définir les règles, plusieurs itérations sont nécessaires. A chaque itération, on divise les individus (ou observations) en k classes (souvent $k = 2$) pour expliquer la variable de sortie.

La première division est obtenue en choisissant la variable explicative qui donne la meilleure séparation des individus. Cette division définit des **sous-populations**, représentées par les **nœuds** de l'arbre.

A chaque **nœud** est associée une **mesure de proportion**.

A la fin des itérations, on obtient des nœuds terminaux, appelés **feuilles** de l'arbre.

Chaque **feuille** est caractérisée par un chemin spécifique à travers l'arbre qu'on appelle une **règle**.

L'ensemble des **règles** pour toutes les feuilles constitue le **modèle**.

L'interprétation d'une règle est aisée si l'on obtient des feuilles pures (100% de variables à expliquer sont VRAI ou FAUX pour une règle donnée. Sinon, il faut se baser sur la distribution empirique de la variable à prédire sur chaque nœud de l'arbre.

Attention : un arbre de décision peut être vite mener à du overfitting. En effet, il peut décrire parfaitement un jeu de données, avec un individu par feuille dans le cas extrême. Dans ce cas, les règles ne sont absolument pas extrapolables et il faut absolument éviter cette situation. Il est donc impératif de s'arrêter à un nombre de feuilles adéquat lors de la réalisation de l'arbre. Dans le jargon, on parle d'**élaguer l'arbre**.

2. CONSTRUCTION DE L'ARBRE DE DECISION

Pour construire l'arbre attendu, il faut répondre à trois questions principales :

1. **Comment choisir la variable de division ?**

   Il faut en effet pouvoir définir l'arborescence de l'arbre en sélectionnant les variables, de la plus discriminante à la moins discriminante.

2. **Comment traiter les variables continues ?**

   Pour définir des nœuds à partir d'une variable continue, l'on doit savoir « couper » la variable continue, pour que ses valeurs inférieures et supérieures à cette coupe puissent caractériser des nœuds distincts.

3. **Comment définir la taille de l'arbre ?**

   L'objectif est de situer le niveau de nœuds optimal, pour trouver le juste équilibre entre sur-apprentissage et arbre trivial.

La réponse à ses questions dépend de l'algorithme utilisé pour construire l'arbre.

3. EXEMPLE

Les figures ci-dessous illustre deux arbres obtenus par l'analyse de l'activité d'un web adviser. L'objectif est d'expliquer le taux de clics sur une bannière publicitaire.

A partir des mêmes données, les deux arbres apportent des informations différentes.

- Dans la figure 1, l'arbre représente un arbre de classification. Il indique si le taux de clics est supérieur ou non à une valeur seuil, pour une règle donnée.
  Par exemple, le chemin le plus à gauche indique que pour les publicités de taille 728x90 et pour un affichage de type BELOW ou UNKNOWN, ce taux n'est supérieur à la valeur seuil que dans moins de 20% des cas. Pour les tailles 300x250 il est supérieur au seuil dans plus de 80%.

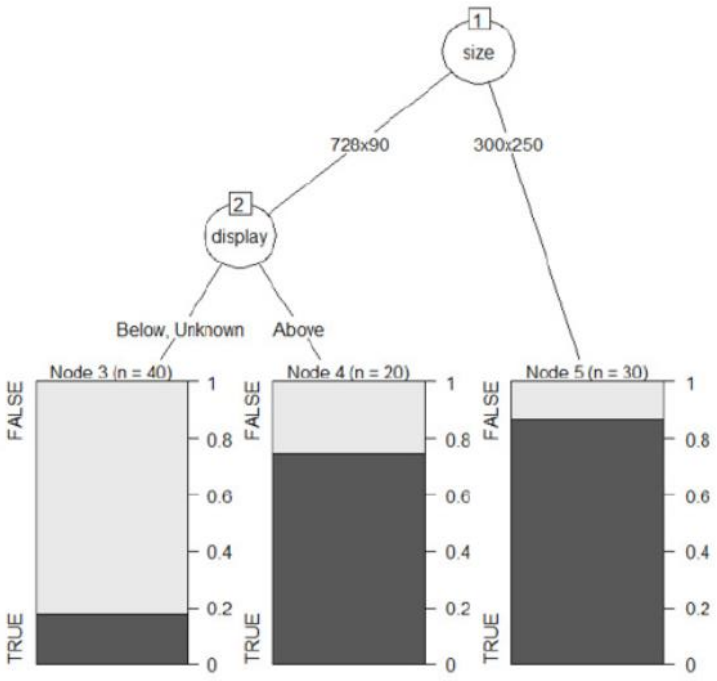

Figure 1 : Arbre de classification

- Dans la figure 2, l'arbre représente un arbre de régression. Chaque feuille utilise une boîte à moustache pour représenter la distribution empirique du taux de clics pour une règle donnée.

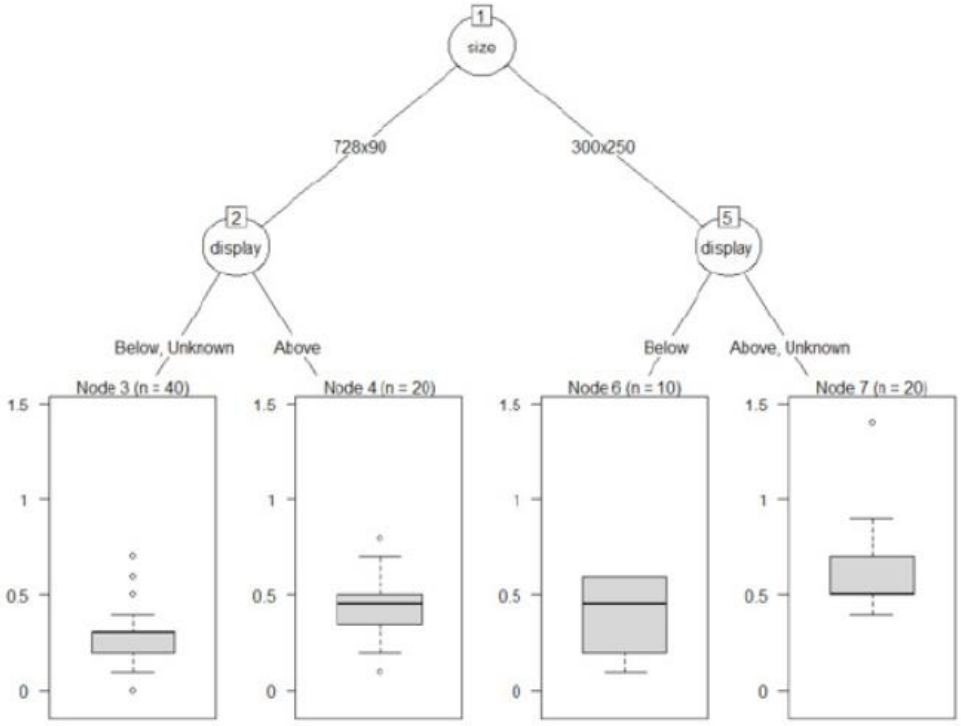

Figure 2: Arbre de régression

4. CONCLUSIONS

- Les arbres de décision sont des modèles d'apprentissage supervisé pour répondre à des problèmes de classification **OU** de régression.
- Un arbre de décision permet de construire des règles à partir des données qui permettent de les ordonner.
- Un arbre de décision est peu gourmand en mémoire.
- Il y a un risque très important d'overfitting. C'est-à-dire que l'arbre n'est pas utilisable car il fournit de bons résultats pour le set d'entrainement, mais utilisé en conditions réelles il classe très mal les nouveaux exemples. Pour pallier à cela, il est possible de diviser le set d'entrainement en deux afin d'avoir un set de données permettant de valider s'il y a overfitting ou pas.
- Un arbre de décision est un moyen puissant de prendre des décisions quand on fait face à un gros volume d'informations

III. POUR LA CULTURE

Classification and Regression Analysis with Decision Trees

Arbres de décision de Ricco Rakotomalala - chercheur