Arbres de Décision

Rindra LUTZ / Nicolas ALLIX

14/01/2021

Qu'est ce qu'un arbre de décision

En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe.
L'ensemble des nœuds se divise en trois catégories :

- Nœud racine : l'accès à l'arbre se fait par ce nœud
- Nœuds internes : les nœuds qui ont des descendants, qui sont à leur tour des nœuds
- Nœuds terminaux : nœuds qui n'ont pas de descendant

Les arbres de décision sont une catégorie d'arbres utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prises dans les nœuds feuille (noeuds terminaux).

Exemple de problème adapté à un approche par arbres de décision : comment répartir une population d'individus (clients, produits, utilisateurs, etc.) en groupes homogènes selon un ensemble de variables descriptives (âge, temps passé sur un site Web, etc.) et en fonction d'un objectif fixé (variable de sortie, par exemple chiffre d'affaires, probabilité de cliquer sur une publicité, etc.).

Il existe 2 types d'arbres de d´ecisions pour la prédiction :

- Les arbres de régressions, qui permettent de prédire une réponse quantitative
- Les arbres de classifications, qui permettent de prédire une réponse qualitative

Pour la petite histoire

La construction des arbres de décision à partir de données est une discipline déjà ancienne. Les statisticiens en attribuent la paternité à Morgan et Sonquist (1963) qui, les premiers, ont utilisé les arbres de régression dans un processus de prédiction et d'explication (AID – Automatic Interaction Detection). Il s'en est suivi toute une famille de méthodes, étendues jusqu'aux problèmes de discrimination et de classement, qui s'appuyaient sur le même paradigme de la représentation par arbres (THAID – Morgan et Messenger, 1973 ; CHAID – Kass, 1980). On considère généralement que cette approche a connu son apogée avec la

méthode CART (Classification and Regression Tree) de Breiman (1984) décrite en détail dans une monographie qui fait encore référence aujourd'hui.

Construire des arbres de décision à partir de données est un sujet ancien. Les statisticiens attribuent cela à Morgan et Sonquist (1963), qui ont été les premiers à utiliser des arbres de régression dans le processus de prédiction et d'interprétation (AID-Automatic Interaction Detection). Puis vint tout un ensemble de méthodes, élargies pour distinguer et classer les problèmes, à partir de représentations arborescentes (THAID-Morgan et Messenger : 1973, ou CHAID-Kass : 1980). Cette méthode est considérée comme ayant atteint son apogée avec la méthode CART de Breiman (1984), qui est décrite en détail dans une monographie et est toujours utilisée aujourd'hui.

Construction d'un arbre de décision

La popularité de la méthode des arbres de décision repose en grande partie sur sa simplicité. Il s'agit de trouver un partitionnement des individus que l'on représente sous la forme d'un arbre de décision. L'objectif est de produire des groupes d'individus les plus homogènes possibles du point de vue de la variable à prédire. Il est d'usage de représenter la distribution empirique de l'attribut à prédire sur chaque nœud de l'arbre.

Contrairement aux arbres de classification, les arbres de régressions sont constitués de variables quantitatives. A chaque noeud, on associe une coupure, et à chaque coupure, une variable de coupure Xjt selon laquelle on découpera le noeud.
Pour qu'un noeud soit pertinent, le seuil de coupure doit avoir une valeur qui maximise un caractère. On note ce seuil de coupure Øt :

- Si $X_i^{jt} \leq \emptyset t$ alors on continue dans la branche de gauche
- Si $X_i^{jt} \leq> \emptyset t$ alors on continue dans la branche de droite

Pureté

On considère un noead pur si tous les individus associés à une des valeurs appartiennent effectivement à cette classe.
La pureté d'un noeud se mesure avec l'indice de Gini, plus la valeur de l'indice est proche de 0, plus le noeud est pur.

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

Coût du noeud

Le coût du noeud va permettre de mesurer la pertinence du choix de la variable de décision. Ce coût est calculé via la formule suivante :

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

Les variables G (gauche et droite) mesure l'impureté des noeuds descendants, respectivement ceux de gauche et ceux de droite.
Les variables m (gauche et droite) sont quant à elles la représentation de la proportion de la population sur les noeuds à gauche et à droite.

Avantages…

- Facilité de mise en oeuvre : un graphe simple à réaliser même si la partie chiffrage nécessite une analyse précise.

- Facilité de prise de décision : modélisation des options possibles, visualisation les différents scénarios.

- Simplification des décisions complexes : le mode graphique permet de comparer de multiples chemins. Chaque nœud peut être mis en perspective et chiffré parmi un ensemble d'hypothèses.

… et limites

Malgré leur simplicité apparente et leur grande utilisation, les arbres de décisions possèdent quelques limites fondamentales, telle que l'instabilité ou le problème du sur-apprentissage. Cette instabilité se traduit par une sensibilité aux fluctuations des échantillons observés. Ainsi, une variabilité quelconque, grande ou petite, peut faire changer nombre de résultats.

Une autre limite des arbres de décision se situe dans le problème du sur-apprentissage. Vous pourrez retrouver une explication de ce qu'est le sur-apprentissage dans mon travail sur la Cross-Validation.

Afin de régler ces problèmes, le Random Forest est une solution adaptée. Cet algorithme va permettre de créer une forêt d'arbres aléatoires issus d'une même base de donnée. Le résultat final à retenir est alors la moyenne de la valeur prédite par chacun des arbres.

On peut également prendre en compte les deux points suivants.

- Le chiffrage reposant sur des estimations, la précision des chiffres joue un rôle primordial dans la pertinence du modèle.

- Un arbre de décision ne prend pas en compte tous les facteurs, notamment ceux reposant sur des évaluations qualitatives. Dans notre exemple : l'entreprise est-elle culturellement prête à se développer à l'international ?

Il n'en demeure pas moins que les variables manquantes émergeront rapidement lors des discussions.

Arbres de classification

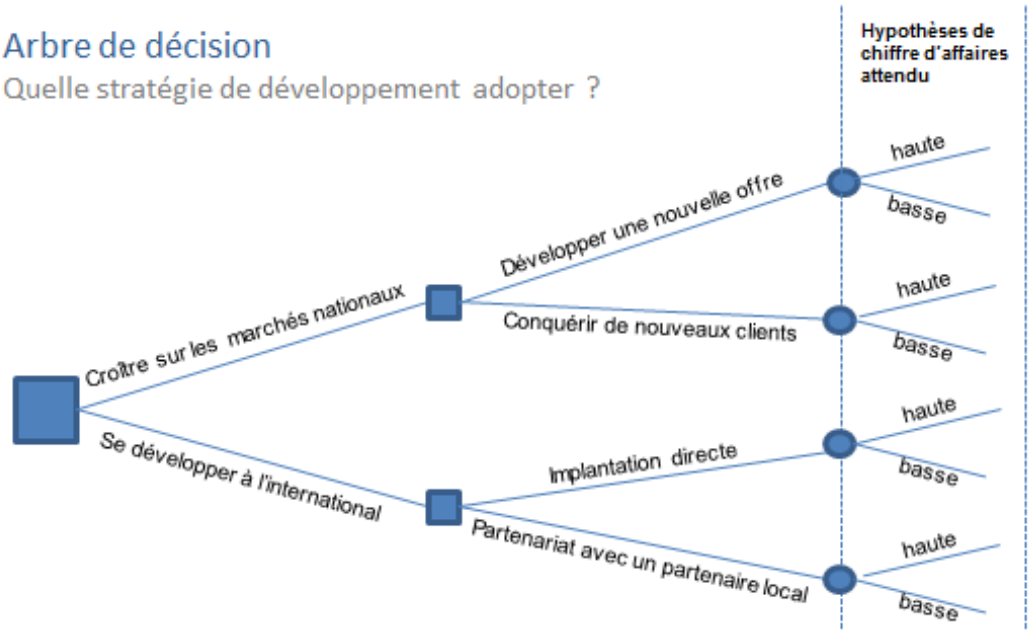
Les noeuds des arbres de classification traitent de caractères qualificatifs. La variable qualitative est notée Xijt et ∈ R.
Suite au test du noeud, l'observation rejoint un groupe parmi A_t, A_t^c:

- Si $X_i^{jt} \in A_t$ alors on continue dans la branche de gauche
- Si $X_i^{jt} \notin A_t$ alors on continue dans la branche de droite

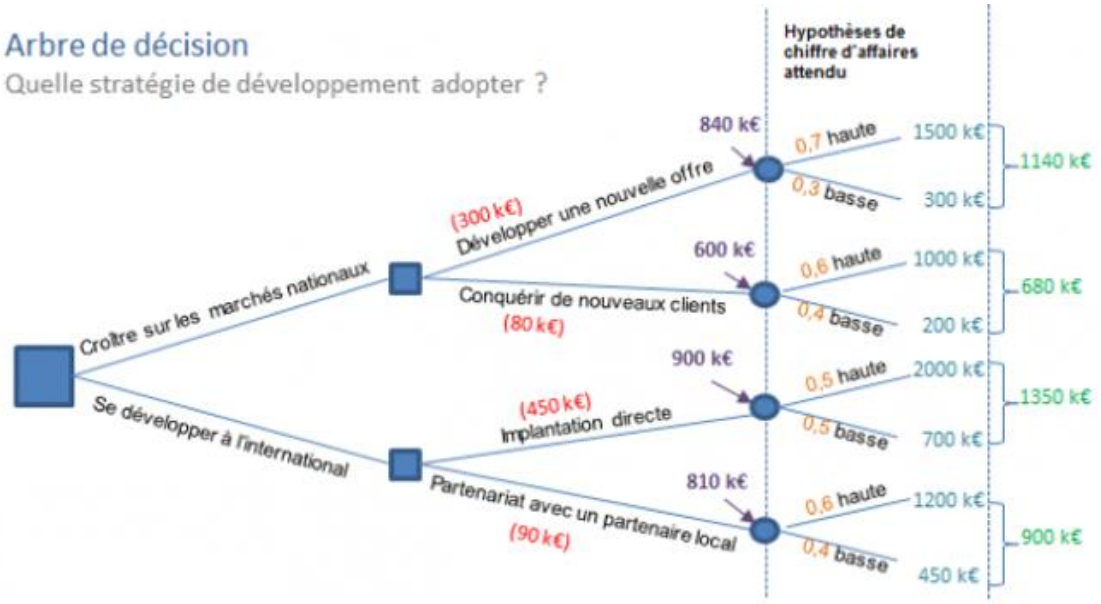
Comme précédemment avec les arbres de régression, il est commun de calculer les indices de Gini et le coûts de chacun des noeuds. Cela permet de voir si les optimums locaux ont été atteints lors de la construction.
Il est nécessaire de porter une attention particulière à cela, car la recherche d'optimum locaux ne conduit pas systématiquement à un arbre optimal. Cela fait partie des limites de la méthode des Arbres de Régression et de Classification.

Arbres de décision : Illustration par l'exemple

Prenons l'exemple d'un comité de direction : le comité doit décider d'élaborer une définition stratégique du chiffre d'affaires de son entreprise. Il peut y avoir plusieurs options, comme se concentrer sur le marché domestique en développant une nouvelle gamme de produits ou en renforçant la manière de trouver de nouveaux clients. Mais une autre option est envisageable : réaliser un développement international par une présence directe ou l'établissement de partenariats locaux. Examinons les arbres de décision que ces décideurs peuvent construire.



Un autre avantage des arbres de décision est qu'ilss simplifiennt le calcul du coût des options. Par conséquent, le choix sera lié au choix qui produit le profit le plus élevé.



Explications

Définition des chiffres

- en orange : la répartition de la probabilité de résultat pour les hypothèses haute et basse
- en bleu : le chiffre d'affaires prévisionnel par hypothèse
- en vert : le chiffre d'affaires prévisionnel par option
- en rouge : le coût de l'option
- en violet : le revenu net prévisionnel de l'option (chiffre d'affaires prévisionnel - coût)

Détails

Prenons le cas du développement d'une nouvelle offre

Hypothèse	Probabilité	Chiffre d'affaires prévisionnel par hypothèse	Chiffre d'affaires prévisionnel pondéré par hypothèse	Chiffre d'affaires prévisionnel par option	Coût de l'option	Revenu net prévisionnel
Haute	0.7	1500 k€	1050 k€ (1500*0.7)	1140 k€ (1050 + 90)	300 k€	840 k€ (1140 - 300)
Basse	0.3	300 k€	90 k€ (300*0.3)			

Dans ce cas d'école, l'implantation directe apparaît comme étant la solution la plus intéressante en terme de revenu, mais il est naturel qu'une telle décision doive prendre en compte d'autres facteurs, tels que le montant de l'investissement, les compétences à mobiliser, etc…