Les arbres de décision
(Apprentissage supervisé)

1. Introduction

Un arbre de décision est un modèle très simple. Etant donnée plusieurs caractéristiques, la décision commence par un de ces caractéristiques; si ce n'ai pas suffisant, on utilise une autre, ainsi de suite. Il est largement connu et utilisé dans de nombreuses entreprises pour faciliter le processus de prise de décision et l'analyse des risques. Il a été largement utilisé dans les années 1960-1980 pour la construction de systèmes experts. Les règles sont introduites manuellement, pour cette raison ce modèle a perdu sa popularité après les années 80. L'apparition des méthodes mathématiques pour construire les arbres de décision fait revenir ce modèle à la bataille des algorithmes de l'apprentissage automatique.

2. Définition

Un arbre de décision est un algorithme d'apprentissage supervisé non paramétrique, qui est utilisé à la fois pour les tâches de classification et régression. Il a une structure hiérarchique, une structure arborescente.

Les arbres de décision sont des classifieurs pour des données représentées par des ensembles attribut/valeur. Un arbre est constitué :

- De nœuds qui testent les attributs
- De branches qui représentent chacune une valeur de l'attribut testé dans le nœud dont elles sont issues
- De feuilles (ce sont les nœuds terminaux de l'arbre) qui indiquent la classe résultante

Le but d'un arbre de décision est de permettre de faire de la prédiction : déterminer la classe d'un nouvel exemple à partir des valeurs de ses attributs.

La prédiction est utilisée par la plupart des sites de vente, lorsque vous allez sur des sites de e-commerce, on vous propose souvent des articles susceptibles de vous intéresser. Pour cela les sites collectent un grand nombre de données sur les achats et les pages visitées par les clients et à l'aide de ces données, ils peuvent ainsi déterminer quel produit ou quelle catégorie de produit est acheté en fonction des caractéristiques des clients. Lorsqu'un nouveau client surfe sur le site, ce dernier est capable de proposer des articles susceptibles de lui plaire en fonction des pages qu'il visite par exemple et ainsi augmenter les chances d'achats.

Il existe plusieurs algorithmes automatiques pour construire les arbres de décision:

- **ID3** (Iterative Dichotomiser 3): développé en 1986 par Ross Quinlan. Il peut être appliqué seulement sur les caractéristiques nominales. Il est utilisé pour le classement.
- **C4.5**: une extension de ID3 par Ross Quinlan. Il peut être appliqué sur tous les types de caractéristiques. Il est utilisé pour le classement.
- **C5.0**: une extension commerciale de C4.5, toujours par Ross Quinlan.

- **CART** (Classification and Regression Trees): comme C4.5 mais utilise d'autres métriques. Aussi, l'algorithme supporte la régression.

L'algorithme général de création d'un arbre de décision:

1. Déterminer la meilleure caractéristique dans l'ensemble de données d'entrainement.
2. Diviser les données d'entrainement en sous-ensembles contenant les valeurs possibles de la meilleure caractéristique.
3. Générez de manière récursive de nouveaux arbres de décision en utilisant les sous-ensembles de données créés.
4. Lorsqu'on ne peut plus classifier les données, on s'arrête.

La construction d'un arbre de décision ne se fait pas au hasard. Le but est de déterminer les meilleurs attributs à placer à chaque nœud pour que l'arbre construit soit le plus petit possible (plus l'arbre est petit plus la prédiction sera facile à obtenir) et qu'en même temps l'arbre soit capable d'effectuer de bonne prédiction. La construction d'un arbre se fait à partir d'un ensemble de données appelé base d'apprentissage. Une fois l'arbre construit, on le teste généralement sur un ensemble de données appelé base de test.

3. Exercice d'application

•Soit à un ensemble de jours. Chaque jour est caractérisé par un numéro et ses conditions météorologiques (***ciel, température***, ***humidité de l'air***, ***force du vent***).
L'attribut cible étant « ***jouer au tennis ? »***, dont les valeurs possibles sont Y = **{oui**, **non}**.
•Une fois l'arbre de décision construit, on pourra classer une nouvelle donnée pour savoir « ***si on joue ou non ce jour-là*** ».

	Ciel	Température	Humidité	Vent	Jouer au tennis ?
1	Ensoleillé	Chaude	Elevée	Faible	Non
2	Ensoleillé	Chaude	Elevée	Fort	Non
3	Couvert	Chaude	Elevée	Faible	Oui
4	Pluie	Tiède	Elevée	Faible	Oui
5	Pluie	Fraîche	Normale	Faible	Oui
6	Pluie	Fraîche	Normale	Fort	Non
7	Couvert	Fraîche	Normale	Faible	Oui
8	Ensoleillé	Tiède	Elevée	Faible	Non
9	Ensoleillé	Fraîche	Normale	Faible	Oui
10	Pluie	Tiède	Normale	Fort	Oui
11	Ensoleillé	Tiède	Normale	Fort	Oui
12	Couvert	Tiède	Elevée	Fort	Oui
13	Couvert	Chaude	Normale	Faible	Oui
14	Pluie	Tiède	Elevée	Fort	Non

On va appliquer l'algorithme ID3 qui se base sur deux concepts l'entropie de Shannon et le gain :

Entropie : $H(X) = -\sum_{i=1}^{n} p_i log_2(p_i)$

Gain : $Gain(X, ai) = H(X) - \sum \frac{|X_{ai=v}|}{|X|} H(X_{ai=v})$

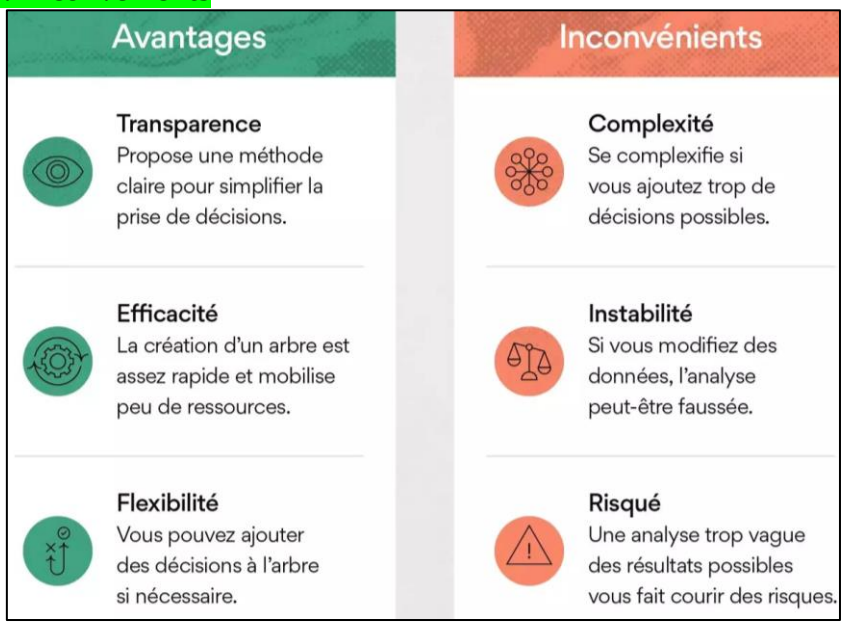
Notre population contient 9 individus qui ont joué au tennis, et 5 n'ayant pas joué alors l'entropie de cette population est :

$$H(X) = -\left(\frac{9}{14}log_2\left(\frac{9}{14}\right) + \frac{5}{14}log_2\left(\frac{5}{14}\right)\right) = 0{,}940$$

4. Application en Python

Exécuter et corriger les notebooks associés

5. Avantages / inconvénients



6. Exercice

Une banque dispose des informations suivantes sur un ensemble de clients :

client	M	A	R	E	I
01	moyen	moyen	village	oui	oui
02	élevé	moyen	bourg	non	non
03	faible	âgé	bourg	non	non
04	faible	moyen	bourg	oui	oui
05	moyen	jeune	ville	oui	oui
06	élevé	âgé	ville	oui	non
07	moyen	âgé	ville	oui	non
08	faible	moyen	village	non	non

L'attribut client indique le numéro du client ; l'attribut M indique la moyenne des crédits sur le compte du client ; l'attribut A donne la tranche d'âge ; l'attribut R décrit la localité du client ; l'attribut E possède la valeur oui si le client possède un niveau d'études supérieur au bac ; l'attribut I (la classe) indique si le client exécute ses opérations de gestion de compte via Internet.

1. Construire l'arbre de décision correspondant à cette base en utilisant l'algorithme ID3.

2. Donner la précision de l'arbre construit sur la base suivante :

client	M	A	R	E	I
01	moyen	âgé	village	oui	oui
02	élevé	jeune	ville	non	oui
03	faible	âgé	village	non	non
04	moyen	moyen	bourg	oui	non