|  |
| --- |
| Université Lille 1 – Master 1 |
| RDF – TP n°09 |
| Arbres de décision |

|  |
| --- |
| BARCHID Sami – SLIMANI Anthony  24/03/2019 |

# Introduction

L’objectif de ce TP est de comprendre le modèle des arbres de décision en utilisant l’algorithme ID3, une variante de l’algorithme général « CART ».

Ce rapport est divisé en TODO parties :

# Question de bon sens

### Problème posé

Deux joueurs (notés A et B) jouent au jeu suivant: A choisit un nombre entier x entre 1 et N (N étant un nombre connu de A et B), et B doit le deviner en lui proposant successivement des valeurs. Pour chaque valeur possible de x proposée par B, A répond “gagné !”, “x est plus petit”, ou “x est plus grand”.

a) Avant de jouer, B dit à A: “je n’ai pas besoin de plus de 4 propositions pour gagner”. A répond alors à B: “C’est vrai, mais avec mon choix de x tu auras besoin d’exactement 4 propositions”. Que vaut N ?

b) B répond alors: “Avec l’information que tu viens de me donner, je suis presque sûr de gagner avant.” Expliquer le raisonnement de B et calculer la quantité d’information que lui a donnée A par sa réponse.

### a) Que vaut N ?

Il faut deviner un numéro entre 1 et N. La manière de trouver ce numéro en posant le moins de questions possible est de toujours choisir un numéro de telle sorte que l’ensemble des numéros plus grands et l’ensemble de numéros plus petits aient tous les deux la même taille. Nous voulons donc maximiser le gain d’information à chaque choix de numéro. En bref, il faut choisir toujours la médiane des numéros possibles.

Si nous utilisons cette méthode optimale en nombre de questions et que nous savons que nous n’avons besoin que de 4 choix maximum, on peut donc conclure que **N=16**.

En effet, en appliquant la stratégie optimale, on est sûr de tomber sur le bon numéro en 4 coups. Par exemple imaginons que A choisisse le numéro à trouver x = 1 :

* B choisi 8, A répond « x est plus petit »
* B choisi 4, A répond « x est plus petit »
* B choisi 2, A réponse « x est plus petit »
* B a trouvé x = 1

Ceci montre bien qu’il est obligatoire, avec cette stratégie, de trouver une valeur unique en 4 étapes maximum.

### b) Expliquer le raisonnement de B et calculer la quantité d’information que lui a donné A par sa réponse.

Par sa réponse, A assure qu’il est impossible que B trouve le numéro en moins de 4 coups. Cela montre que le numéro que A a choisi est impair. En effet, la stratégie de B implique de devoir toujours choisir un numéro pair si nous pouvons séparer l’ensemble des numéros.

Le seul moment où B choisirait un numéro impair serait au moment où il n’y a plus que ce choix-là de disponibles (et donc que B trouve le bon numéro en 4 coups). A a donc donné l’information que le numéro choisi est **impair**.

B n’a plus qu’à travailler uniquement sur les nombres impairs et donc est presque sûr de pouvoir donner sa réponse en moins de 4 coups. Il n’a qu’à utiliser la même stratégie qu’auparavant mais sur l’ensemble des numéros impairs entre 1 et N.

# Variante du jeu du pendu

L’exercice proposé pour ce TP que nous étudierons afin de comprendre les arbres de décision est une variante du jeu du pendu.

Le jeu se déroule de la manière suivante :

* Un ensemble de noms possibles est mis à disposition
* Le joueur en choisi un
* Tant que l’ordinateur n’a pas trouvé le nom choisi
  + Il pose une question du type : « Ce mot contient-il la lettre xxxxxx ? »
  + Le joueur répond Oui ou non

Le but étant que l’ordinateur puisse trouver le mot le plus rapidement possible.

Lorsque nous faisons le lien avec la théorie des arbres de décisions, nous remarquons :

* La variable à prédire (la classe) sont les mots eux-mêmes.
* Les variables prédictives sont les lettres de l’alphabet qui composent le mot.