MASTER MIAGE 2ème année
Université Paris Nanterre

Mémoire de fin d'études

Optimisation de parcours d'arbre de décisions dans le cadre du MNIST

Université Paris Nanterre

Auteur :
Alexandre PETIT-PAS

Tuteur :
P. Marie-Pierre GERVAIS

Résumé

Résumé

Remerciements

Remerciements

Sommaire

Introduction

Motivations

Objectifs

Chapitre 1

Concepts formels

1.1 Introduction

L'objectif de ce chapitre est dans un premier temps de rappeler des notions de probabilités [1, 2, 3] qui serviront à la compréhension de ce document. Il s'agit ensuite d'expliquer dans sa généralité les concepts d'arbres décisionnels afin d'introduire des principes de classement et d'optimisation.

1.2 Rappels de probabilité

1.2.1 Expérience aléatoire

Une expérience est dire aléatoire si les résultats possibles sont connus à l'avance sans vraiment savoir celui obtenu au préalable. On appelle univers, noté Ω, l'ensemble de toutes les issues possibles d'une expérience et ω une réalisation de l'expérience.

$$\Omega = \left\{\omega_1, \omega_2, \cdots, \omega_n\right\}, n \in \mathbb{N} \tag{1.1}$$

1.2.2 Événements

Un événement correspond à un ensemble de résultats possibles pour une expérience. Si cet ensemble est constitué d'un seul élément, on parle alors d'événement élémentaire. Si l'ensemble de résultats est égal à l'univers Ω, alors l'événement est dit certain. En revanche, si aucun résultat n'est présent (ensemble $\emptyset$), alors c'est un événement impossible.

1.2.3 Probabilités

Une probabilité est une fonction qui à un événement A, associe un poids.

$$\begin{cases} P(\Omega) = 1 \\ P(\emptyset) = 0 \\ 0 \leq P(A) \leq 1 \\ P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \end{cases} \tag{1.2}$$

Plus la probabilité est proche de 1, plus il est possible que l'événement se réalise. Soit A et B deux événements quelconques, $P(A)$ est dite conditionnelle si son résultat est influencé par l'événement B :

$$\begin{cases} B \neq \emptyset \\ P(B) \neq 0 \\ A = (A \cap B) \cup (A \cap \overline{B}) \end{cases} \tag{1.3}$$

On peut alors déduire deux formules : le théorème de Bayes (1.4) permettant de calculer $P_B(A)$ et le théorème es probabilités totales (1.5) qui permet de connaitre la valeur de $P(A)$ à partir de deux événements A et B.

$$P_B(A) = \frac{P(A \cap B)}{P(B)} \tag{1.4}$$

$$P(A) = P_B(A)P(B) + P_{\overline{B}}(A)P(\overline{B}) \tag{1.5}$$

1.3 Arbre de décision

1.3.1 Définition

Un arbre de décision est une représentation structurelle qui permet d'aboutir à un choix. C'est un graphe acyclique orienté composé :

- d'un sommet sans parents appelé racine,
- de sommets appelés nœuds, correspondant à des tests,
- de sommets terminaux nommés feuilles,
- des arrêtes, ou branches, désignant chacune les résultats d'un test

Pour construire un arbre de décision, une approche *Top-Down* est utilisée, appelée *Top Down Induction of Decision Tree* (*TDIDT*). Elle peut se décomposer en plusieurs parties :

1. Partir du jeu complet de données et construire la racine.
2. Réaliser un test afin de séparer les données.
3. Séparer le nœud actuel en fonction des résultats possibles.
4. Appliquer récursivement jusqu'à atteindre les feuilles.

Lorsque chaque nœuds sont composés exactement de deux descendants (hors feuilles), on parle alors d'arbre de décision binaire. C'est un cas très largement utilisé des algorithmes de construction d'arbres tels que CART [4], qui sera expliqué ultérieurement.

1.3.2 Exemple

Prenons comme exemple les données météorologiques présente dans la table 1. Cette table est constitué de différentes colonnes concernant différentes information ainsi qu'une colonne indiquant si la décision de sortir a été prise ou non. Dans un peu premier temps, il est possible de déduire les tests à réaliser comme par exemple ”La température est-elle élevée ?" ou encore "Quel est le temps ?". A ces questions découlent des réponses possibles, {oui, non} pour la première et {soleil, nuageux, pluie} pour la deuxième. Une fois l'approche *TDIDT* utilisée, un arbre de décision est alors obtenu (figure 1.1).

Jour	Temps	Température	Humidité	Vent	Sortie
1	soleil	élevée	haute	faible	N
2	soleil	élevée	haute	fort	N
3	nuageux	élevée	haute	faible	Y
4	pluie	moyenne	haute	faible	Y
5	pluie	basse	normale	faible	Y
6	pluie	basse	normale	fort	N
7	nuageux	basse	normale	fort	Y
8	soleil	moyenne	haute	faible	N
9	soleil	basse	normale	faibles	Y
10	pluie	moyenne	normale	faible	Y
11	soleil	moyenne	normale	fort	Y
12	nuageux	moyenne	haute	fort	Y
13	nuageux	élevée	normale	faible	Y
14	pluie	moyenne	haute	fort	N

Table 1.1 – Table Météo

1.4 Arbre de décision pour le classement

1.4.1 Attributs et classes

Soit un ensemble d'observation S de taille n :

$$S = \{s_1, s_2, ..., s_n\}. \tag{1.6}$$

Chaque observation s_i est composée d'un ensemble de m attributs :

$$A_s = \{a_1, a_2, ..., a_m\}. \tag{1.7}$$

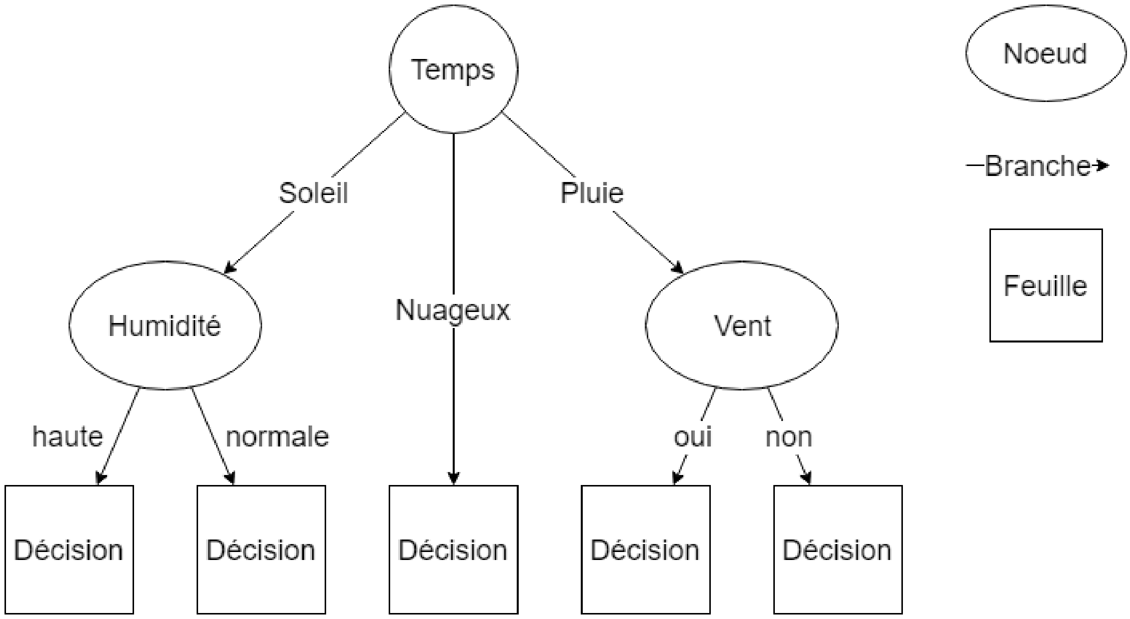

Figure 1.1 – Arbre de décision de la table Météo

Soit V_j l'ensemble de taille l des valeurs possibles de l'attribut a_j d'une observation s_i tels que :

$$V_j = \{v_1, v_2, ..., v_l\}. \tag{1.8}$$

La valeur de l'attribut a_j sera représentée par v_k. Un attribut est dit qualitatif si l'ensemble des valeurs possibles est symbolique (non numérique), par exemple si V_j représente les couleurs d'écriture d'un mot. On obtiendrait alors $V_j = \{$ bleu, rouge, noir, vert $\}$.

Une donnée est quantitative si l'ensemble des valeurs possibles est un ensemble numérique fini ou infini :

- Si un attribut peut prendre une infinité de valeurs dans son ensemble, alors celui-ci est qualifié de continu, par exemple le temps d'exécution d'un processus.
- Dans le cas contraire, une variable dite discrète possède une valeur finie. Elle est généralement liée à une énumération, comme par exemple le nombre de trait dans un caractère.

a_j est nominal si la notion d'ordre n'est pas présente dans l'ensemble des valeurs possibles, par exemple si $V_j = \{$ un, deux, trois, quatre, cinq, six, sept, huit, neuf $\}$ représente le nom d'un chiffre en toutes lettres.

Un attribut est ordinal si la notion les valeurs possibles contiennent la notion d'ordre. Cela peut être par exemple l'appréciation d'un client : $V_j = \{$ mauvais, bon, très bon $\}$

Une variable est qualifiée de binaire si l'ensemble V_j des valeurs possibles est de taille $l = 2$

La classe d'une observation correspond a une "catégorie" et permet de se rapprocher ou de se différencier des autres observations. Elle correspond à une feuille dans un arbre décisionnel.

Reprenons la table 1.1 (Météo), l'attribut jour correspond à l'identifiant dans l'observation et donc n'est pas pris en compte. *Temps*, *Température*, *Humidité* et *Vent* sont des attributs qualitatifs dont deux d'entre eux binaires : $V_{Humidite} = \{haute, normale\}$ et $V_{Vent} = \{fort, faible\}$. Chaque observation possède une classe *Sortie* pouvant prendre Y ou N comme valeur.

1.4.2 Apprentissage supervisé

Le domaine de l'apprentissage peut se séparer en deux types : le supervisé et le non-supervisé.

Dans le cas de l'apprentissage non-supervisé, le but recherché est d'assembler les similitudes entre les observations. Une des méthodes les plus communes est le *clustering* [5].

Le second type d'apprentissage est dit supervisé. Dans ce cas, le but est d'apprendre du modèle pour arriver à déterminer la classe de l'observation. C'est dans ce domaine que se situe les arbres de décision. Les jeux de données sont alors séparés en deux : une partie jeu d'entrainement (*training set*) et une partie jeu de test (*test set*). Le but est d'apprendre du jeu d'entrainement afin de classer les observations du jeu de test. Soit C un ensemble de classe de taille n :

$$C = \{c_1, \cdots, c_n\} \tag{1.9}$$

Classer une observation s_i revient à trouver le c_j qui lui correspond le mieux par le biais d'une fonction de classement F, calculé grâce au jeu de données d'entrainement.

1.4.3 Quantification de l'information

La quantification de l'information permet de dégager une composante importante des tests d'un arbre de décision : le manque d'information. *Dans la théorie de l'information, les notions de quantité d'information et d'incertitude sont équivalentes* [2, 6]. Partant de ces concepts, la quantité d'information h(x) d'une probabilité P(x) est une fonction croissante : si P(x) augmente, h(x) augmente. Par ailleurs, les événements certains et impossibles n'apportent aucune information étant donné que le résultat est connu d'avance.

Afin de mesurer le gain d'information, il est nécessaire d'introduire le concept d'entropie correspondant à l'impureté d'une observation en théorie de l'information [7].

Soit S un ensemble d'observations de taille n où $s_i = \{Y, N\}$. L'entropie E de S, où p_Y correspond à la probabilité d'avoir Y et p_N d'avoir N [8], est définie par

$$E(S) = -p_Y log_2 p_Y - p_N log_2 p_N \tag{1.10}$$

La figure 1.2 représente la fonction d'entropie de S en fonction de p_Y. On observe que l'entropie est comprise entre 0 et 1 et vaut 0 lorsque p_Y vaut 0 ou 1. Cela confirme le fait que les événements impossibles et certains n'apportent aucun gain d'information.

Nous venons de montrer l'entropie de S pour un attribut binaire. Par extension, si une observation s_i peut prendre l valeurs, alors

$$E(S) = \sum_{i=1}^{l} -p_i log_2 p_i \tag{1.11}$$

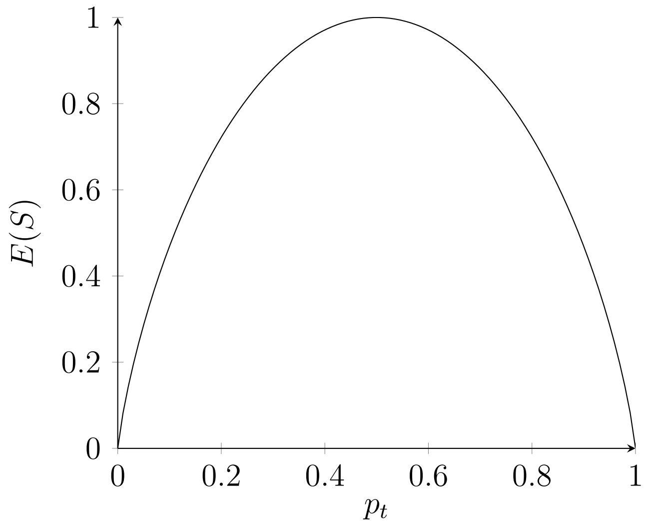

Figure 1.2 – Entropie de S en fonction de p_t

Par exemple, d'après la table 1.1, l'entropie de l'attribut sortie est égale à

$$\begin{aligned} E(Sortie) &= -\frac{9}{14} log_2(\frac{9}{14}) - \frac{5}{14} log_2(\frac{5}{14}) \\ &= 0.4098 - 0.5305 \\ &= 0.940 \end{aligned} \tag{1.12}$$

entropie base 2 entropie générique graphe entropie example sur la table meteo

1.5 Arbre de décision pour l'optimisation

quand arrêter un arbre elagage optimisation d'algorithmes (boosted etc)

Chapitre 2

Lazy Decision Tree

2.1 Lazy Decision Tree pour le classement

2.2 Lazy Decision Tree pour l'optimisation

Chapitre 3

Classement supervisé

3.1 K Nearest Neighbors

3.2 CART

3.3 C4.5

3.4 A *

Chapitre 4

Apprentissage du MNIST

4.1 Définition du MNIST

4.2 Méthode de recherche

4.3 Algorithme de classement

4.4 Algorithme Lazy

4.5 Comparaison des résultats

Bibliographie

[1] Hélène Guérin. *Qu'est ce qu'une probabilité ?* 2008.

[2] Lamis Hawarah. "Une approche probabiliste pour le classement d'objets incomplètemet connus dans un arbre de décision". Thèse de doct. Université Joseph Fourier - Grenoble I, 2008.

[3] R. Gilleron F. Denis. *Apprentissage à partir d'exemples*. 2000.

[4] Jan Kozak. *Decision tree and ensemble learning based on ant colony optimization*. Springer, 2019.

[5] Daniel T. Larose. *Discovering knowledge in data : an introduction to data mining*. Wiley-Interscience, 2005.

[6] Steven Roman. *Introduction to coding and information theory*. Undergraduate texts in mathematics. 1997.

[7] C. E. Shannon. "A Mathematical Theory of Communication". In : *The Bell System Technical Journal, Vol. 27*. 1948, p. 379-423, 623-656.

[8] Tom M. Mitchell. *Machine Learning, International Edition*. McGraw-Hill Series in Computer Science. 1997.

Table des matières

Table des figures

Liste des tableaux