

Rapport de stage
Algorithmes de minimisation du regret et
integrales de choquet
LABORATOIRE DE RECHERCHE EN
INFORMATIQUE

Quentin LIEUMONT

18 Juin 2019

Table des matières

1	Introduction	2
1.1	Objectif du stage	2
1.2	Présentation générale du LRI	2
1.3	Unité GALAC	2
2	Contexte et motivations	3
2.1	Objectif	3
2.2	Un petit point théorique...	3
2.2.1	Réseaux de neurones	3
2.2.2	Intégrales de choquet	7
3	Matériel et méthodes	8
4	Résultats	9
5	Discussion et conclusion	10
	Bibliographie	11

Sous la direction de : Johanne COHEN
Martine THOMAS

Chapitre 1

Introduction

1.1 Objectif du stage

Mon stage c'est déroulé au laboratoire de recherche en informatique, le sujet de mon stage était le suivant : *Algorithmes de minimisation du regret et intégrales de Choquet*. Je n'avais jamais fait d'informatique théorique (excepté IA) et n'ayant jamais vus la théorie de la mesure, il m'était compliqué de comprendre la notion d'intégrale de Choquet (découlant des intégrales de Lebesgue). Ainsi, durant ce rapport, la partie mathématique théorique ne sera pas vue en détail.

1.2 Présentation générale du LRI

Le Laboratoire de Recherche en Informatique (LRI) est une unité de recherche de l'Université Paris-Sud et du CNRS. Créé il y a plus de 35 ans, le laboratoire est localisé sur le plateau du Moulon depuis début 2013. Il accueille plus de 250 personnes dont environ un tiers de doctorants.

Organisés en neuf équipes, les recherches du laboratoire incluent à la fois des aspects théoriques et appliqués (ex : algorithmique, réseaux et bases de données, graphes, bioinformatique, interaction homme-machine, etc). De part cette diversité, le laboratoire favorise les recherches aux frontières de différents domaines, là où le potentiel d'innovation est le plus grand. En plus de son activité de publication (2000 publications entre début 2008 et mi 2013), le LRI développe de nombreux logiciels.

Pour plus d'informations, je vous invite à aller regarder sur leur site où une présentation détaillée est continuellement tenue à jour [1].

1.3 Unité GALAC

Durant mon stage, j'ai travaillé sous la direction de Johanne COHEN qui est la responsable de l'équipe GALAC.

L'équipe GALAC est une équipe du LRI qui travaille sur des thématiques théoriques comme l'algorithmique, la théorie des graphes ou les systèmes en réseaux.

Chapitre 2

Contexte et motivations

2.1 Objectif

Le but du stage était de prédire le juste prix d'une maison à partir de quelques unes de ces caractéristiques. On pourrait tout simplement multiplier le prix au mètre carré par la superficie de la maison mais on verra par la suite que cette technique est loin d'être optimale. Un réseau de neurones et les intégrales de choquets ont donc été utilisées.

2.2 Un petit point théorique...

Afin de mieux comprendre le déroulé du stage et les interactions des différentes notions, un petit point théorique est nécessaire :

2.2.1 Réseaux de neurones

Qu'est ce que c'est ?

On entend souvent dans les médias parler d'intelligence artificielle et de réseaux de neurones, ces deux notions sont radicalement différentes mais elles sont souvent mélangées et confondues sur la place publique...

L'intelligence artificielle est un concept informatique, un paradigme de programmation, ce n'est pas ce que j'ai fait durant mon stage, on n'en parlera pas ici mais si vous voulez en savoir plus, je vous redirige vers la chaîne youtube de Lê NGUYỄN HOANG, chercheur en mathématique qui saura vous l'expliquer bien mieux que moi (à regarder sans moderation) [2].

Un réseau de neurones est une architecture informatique permettant de faire des régressions de fonctions, de la généralisation et de l'optimisation. Il est composé, comme son nom l'indique, de neurones mis en réseau grâce à des connexions.

Un neurone

Un neurone (représenté par un cercle dans les schemas) est une unité de base du réseau, il se decompose en trois phases

L'entrée : Un neurone prend des informations en entrée, de manière générale un nombre réel entre 0 et 1 mais d'autres objets sont envisageables (image, pixel, son...). Il n'y a pas de nombre minimum ou maximums (on pourrait imaginer un neurone ne prenant pas d'entrée, ou au contraire en prenant une infinité), ni de contrainte sur les différents objets.

Exemples :

- un neurone prenant les trois valeurs de couleur d'un pixel (entier entre 0 et 255).
- un neurone prenant les trois valeurs de couleur d'un pixel (réel $[0, 1]$).
- un neurone prenant une séquence ADN en entrée

On note le vecteur entrée X et chacun de ses éléments $x_1 \dots x_i \dots x_n$.

La fonction interne : Le réseau va alors faire un calcul à partir des entrées et de poids, des valeurs définies pour chaque neurone qui peuvent varier durant l'apprentissage.

Le vecteur poids se note W et chacun de ses éléments $w_1 \dots w_i \dots w_n$.

Exemples :

$$\begin{aligned} f(X) &= W.X = \sum_{i=1}^n w_i \times x_i \\ f(X) &= e^{w_1 + x_1} \times w_2 \\ f(X) &= \max(X) \\ f(X) &= \text{"nombre de A sur les } w_1 \text{ premières bases de } x_1 \text{"} \\ &\quad (\text{avec } x_1 \text{ une séquence ADN}) \end{aligned}$$

La fonction d'activation : Comme vu précédemment, ces réseaux sont mis en réseau, il est donc intéressant d'avoir une norme pour l'entrée et la sortie afin de pouvoir lier des neurones entre eux sans distinctions.

La norme qui a été choisie est, comme précédemment cité, un réel entre 0 et 1.

Or, il peut être remarqué que les fonctions ci-dessus ne renvoient pas forcément des nombres entre 0 et 1... On utilise donc la fonction d'activation afin de normaliser les sorties.

Exemples :

- Pour le cas précédent sur l'ADN : $f(X)/w_1$
- Pour une fonction de \mathbb{R} dans \mathbb{R} : fonction sigmoïde
- Pour une fonction de \mathbb{R} dans $[0, 1]$: fonction identité

Pour resumer : On prend *des entrées*, on les passe dans la *fonction principale* du neurone, puis le résultat dans la *fonction d'activation*.

Le résultat ressemble a cela :

$$f_{act}(f(X)) \quad (2.1)$$

Le tout dépendant bien évidemment de W que l'on peut faire varier afin de modifier la fonction du neurone.

Exemples :

Prenons un neurone avec deux entrées : x_1 et x_2 ,
de fonction principale $f(X) = X.W$
et de fonction d'activation identité.
Si on a $W = (1, 1)$, ce neurone fait une somme.
Mais si $W = (0.5, 0.5)$ ce neurone fait une moyenne.

Un réseau

Une fois que nous avons de nombreux neurones faisant chacun des actions bien spécifiques, on voudrais les relier afin de gérer des comportements plus complexes comme faire une somme de moyenne (ou contrôler un drone, prédire les mouvements boursiers...).

Il suffit alors de relier les neurones entre eux, de manière générale de manière linéaire de gauche à droite. De nombreuses manières de relier les neurones existent (recurrent, convolution...), ici on ne parlera que de la connexion dite "Dense" ou "fully connected" : le réseau se découpe en n couches de neurones, chacune composée de neurones dont les entrées sont les sorties des neurones de la couche précédente.

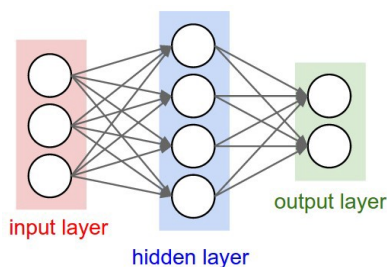


FIGURE 2.1 – Réseau dense simple

Comme on peut le voir sur la figure, les neurones les plus à gauche sont les neurones d'entrée (capteurs) et ceux les plus à droite ceux de sortie (contrôle des moteurs, affichage d'une note...). Tous les autres neurones sont "cachés", ils n'interagissent pas directement avec l'environnement.

L'apprentissage

Jusqu'ici on a vu comment créer une architecture modulable permettant d'approximer toutes les fonctions. Mais une question se pose toujours : Quel est l'intérêt ? Pourquoi ne pas directement écrire la fonction "en dur" ?

La réponse tient en trois mots : *Descente de gradient stochastique* (ou SGD en anglais).

Ce concept est ce qui fait que les réseaux de neurones sont les architectures favorites des data scientists et des chercheurs en IA.

L'idée est assez simple, on recherche une fonction générale dont on ne connaît que certaines valeurs discrètes. On va commencer par définir une fonction nommée "loss function" qui décrit la précision du réseau actuel : elle prend en entrée deux valeurs : la valeur théorique et la valeur obtenue. elle retourne un réel positif, plus il est faible, plus le réseau est proche de la fonction théorique.

Exemples :

$$loss(exp, obt) = |exp - obt|$$

$$loss(exp, obt) = (exp - obt)^2$$

Le problème de régression se résume alors à minimiser la fonction de perte, c'est ici que la descente de gradient stochastique fait son entrée : Pour minimiser la perte, on aimerait bien descendre sa pente, c'est-à-dire dériver suivant les différentes variables¹. Étant donné que de manière générale, il y a plusieurs variables, la dérivée se transforme en gradient. Ainsi, afin de pouvoir calculer facilement les gradients, des applications linéaires sont privilégiées dans les fonctions des neurones.

On a donc expliqué d'où vient la descente de gradient, mais que veut dire stochastique ? Le mot stochastique est synonyme de hasard. En effet le temps de calcul du gradient est exponentiel alors lorsque le réseau est grand (un réseau peut facilement atteindre le million de neurones voir même des milliards [3]) il est inimaginable de calculer la totalité du gradient, on utilise alors le gradient stochastique.

Cette descente de gradient est extrêmement rapide, ainsi, il est possible de l'itérer de nombreuses fois pour essayer de minimiser la fonction de perte. L'espace des solutions étant rarement idéal (rarement une "cuvette" mais constitué de "bosses" et de "trous" chaotiques) il n'est pas obligatoire d'arriver à atteindre le minimum global mais au moins un minimum local sera atteint. Ainsi, avec plusieurs apprentissages on peut approximer très efficacement quasiment toutes les fonctions.

1. NB : Ici les variables sur lesquels on intervient sont les w_i (pas les x_i).

2.2.2 Intégrales de choquet

L'intégrale de choquet est une intégrale découlant de la théorie de la mesure que je n'ai pas vus, de plus, ici on ne parlera que du modèle discret.

Pour l'expliquer simplement, prenons un exemple : Supposons que l'on veuille conseiller des personnes sur l'achat d'un ordinateur. Ces personnes ne connaissent absolument rien en informatique. La seule chose qu'ils veulent c'est une note, plus elle est élevée, plus l'ordinateur est performant. Essayons de faire un algorithme assez simple pour résoudre ce problème : Chaque composant se voit attribué une note (en fonction de la puissance, la qualité de fabrication...), on appelle cette note *l'utilité* du composant. Une fois toutes les utilités étudiées, on donne un poids à chacun des composants. Enfin on fait la somme de utilité fois poids pour tous les composants.

Ainsi, si un ordinateur a de meilleurs composants, plus de puissance, etc., sa note sera supérieure.

Ce modèle paraît raisonnable dans la plupart des cas mais il est extrêmement mauvais dans les cas extrêmes : Supposons qu'un constructeur peu scrupuleux propose un ordinateur assez étrange : Le processeur le moins cher du marché (assez mauvais) mais énormément de RAM, par exemple 128Go. Votre classement le placera forcément en haut de la liste, même si vous en conviendrez, cet ordinateur est assez inutile...

Il faudrait donc trouver un autre système modélisant les interactions entre les composants. Ce modèle est exactement celui sur lequel j'ai travaillé durant mon stage : *les intégrales de choquet*. En plus de donner un poids aux utilités, un poids est donné aux interactions deux à deux des utilités par le biais des fonctions min et max.

Voici donc la fonction que l'on va vouloir approximer :

$$C(X) = \sum_{i=1}^n w_i \times x_i + \sum_{i=1}^n \sum_{j=i+1}^n \left(w_{M\,ij} \times \max(x_i, x_j) + w_{m\,ij} \times \min(x_i, x_j) \right) \quad (2.2)$$

Avec X le vecteur des utilités et W , W_m et W_M les vecteurs des poids.

Chapitre 3

Matériel et méthodes

Chapitre 4

Résultats

Chapitre 5

Discution et conclusion

Bibliographie

- [1] “Laboratoire de recherche en informatique.” www.lri.fr.
- [2] “Science4all.” www.youtube.com/playlist?list=PLtzmb84AoqRTl0m1b82gVLcGU38miqdrC.
- [3] “Spectrum ieee.” spectrum.ieee.org/tech-talk/computing/software/biggest-neural-network-ever-pushes-ai-deep-learning.