

UNIVERSITÉ DE MATHÉMATIQUES AIX-MARSEILLE
ET
DREAMQUARK

Etudes comparatives des méthodes
algorithmiques d'hyperoptimisation

MASTER D'INGÉNIEURIE MATHÉMATIQUES ET MODÉLISATION
ANNÉE 2019/2020

Auteur :
Elie LAGACHE

Responsable de master :
Anne NOURI

Responsable du stage :
Sebastian MUELLER

Responsable professionnel
Alexandre CAMERON

“The human condition is plagued with a labyrinth of shortcomings, frailties and limitations that hinder man from reaching his fullest potential. Therefore, it only makes sense that we find ourselves at the next phase in human evolution where restricted man merges with the infinite possibilities of hyper-evolving technologies. This techno-human transmutation will prove to be ‘the’ quantum leap in human progression. The harmonization of technologically extending oneself, consciousness, artificial intelligence and machine learning will reverse the failures of genetic predisposition and limitation. ” James Scott

Remerciements

Je tiens à remercier Monsieur Sebastian Mueller, Maitre de conférence à l'Université de d'Aix-Marseille, pour sa disponibilité et son encadrement pédagogique.

Je remercie, de plus, Alexandre Cameron, Data Scientist Chief de l'équipe project au sein de Dreamquark, avec qui j'ai réalisé la majeure partie du projet de mon mémoire et qui m'a accompagné pendant toute la durée de mon stage.

Je n'oublie pas de remercier ma responsable de master Mme Anne Nouri qui pu faire en sorte que je puisse finaliser et passer cet UE mémoire dans de bonne condition. Je remercie enfin, l'équipe project de Dreamquark avec qui j'ai participé à plusieurs études intéressantes en lien avec le sujet du mémoire.

Table des matières

1	Introduction	1
1.1	Machine learning	1
1.2	Présentation du stage	1
1.2.1	Environnement professionnel	1
1.2.2	Objectifs et sujets principaux	1
1.3	Introduction du sujet	2
1.3.1	Classification binaire	2
1.3.2	Arbre de décision	3
1.3.2.1	Principe	3
1.3.2.2	Définitions	3
1.3.2.3	Construction	3
1.3.3	Etude d'un cas pratique	4
1.3.3.1	Echantillonnage d'un jeu de données	4
1.3.3.2	Evaluation empirique d'un modèle	4
1.3.3.3	Application	5
1.3.4	Notion d'hyperparamètre	7
1.3.4.1	Paramètre ou hyperparamètre?	7
1.3.4.2	Profondeur de l'arbre : max depth	7
1.3.4.3	Echantillon minimal : min samples split	8
1.3.4.4	Nombre de variables : max features	8
2	Un problème d'optimisation	10
2.1	Notions importantes	10
2.1.1	Echantillon de validation	10
2.1.2	Forêt aléatoire	10
2.2	Optimisation	11
2.2.1	Espaces d'hyperparamètres	11
2.2.2	Equation et objectif	12
2.2.2.1	Problème formel	12
2.2.2.2	Validation croisée	12
2.2.2.3	Fonction-objectif	14
2.3	Grid Search	14
2.3.1	Principe	14
2.3.2	Lancement et analyse d'une grid search	15
2.3.2.1	Grid search : arbre de décision	15
2.3.2.2	Grid search : Forêt aléatoire	16
3	Optimisation bayésienne par processus Gaussien	18
3.1	Une méthode bayésienne	18
3.1.1	Une approche adaptée	18
3.1.2	Principe	19
3.2	Processus Gaussiens	19
3.2.1	Définitions et rappels	20
3.2.2	Distribution de fonctions	21

3.3	Fonction d'acquisition	21
3.3.1	Expected improvement	22
3.4	Application pratique	24
3.4.1	Algorithme final	24
3.4.2	Résultats et analyse	24
4	TPE et recherche aléatoire	27
4.1	Tree-structured Parzen Estimator	27
4.1.1	Estimation par noyau	27
4.1.2	Principe de la methode TPE	28
4.1.3	Optimisation de l'EI	28
4.2	Lancement numérique et analyse	28
4.2.1	Algorithme complet	28
4.2.2	Résultats et analyse	29
4.3	La recherche aléatoire	29
4.3.1	Principe d'une RS	30
4.3.2	Lancement numérique et résultats	31
4.4	Classement des performances	33
4.4.1	Arbre décision	34
4.4.1.1	Maximum d'AUC	34
4.4.1.2	Temps d'entraînement	35
4.4.2	Forêt aléatoire	35
4.4.2.1	Maximum d'AUC	35
4.4.2.2	Temps d'entraînement	35
4.4.2.3	Synthèse	35
4.5	Convergence	37
4.5.1	Résultats et analyses	37
4.6	Limites de l'étude	38
4.7	Résultats et conséquences	38

Chapitre 1

Introduction

Nous introduirons ce mémoire présentant tous les points essentiels du stage réalisé ainsi qu'un exemple introductif du sujet traité.

1.1 Machine learning

Le machine learning (apprentissage automatique ou statistique), représente une étude de l'intelligence artificielle. Il s'agit de laisser évoluer un processus algorithmique par l'intermédiaire d'une "machine" afin de résoudre une problématique statistique (le plus souvent) difficile. L'analyse réalisée est généralement associée à des arbres ou des courbes (par exemple la fonction de perte) et les problématiques à de la classification (binaire ou multi-classes). Il s'agit de l'étiquetage des données à une certaine classe. Un modèle de regression logistique en est un exemple classique.

1.2 Présentation du stage

1.2.1 Environnement professionnel

L'entreprise d'accueil est spécialisée dans l'intelligence artificielle concernant des problématiques de finance et d'assurance. En effet, Dreamquark développe une machine virtuelle d'analyse prédictive utilisable par des personnes métiers (banques, assurances, marketing,..) sans qu'il soit nécessaire de posséder forcément des connaissances approfondies en data science. Les calculs sous-jacents découlent surtout des algorithmes de deep learning (apprentissage profond : réseaux de neurones). Le machine learning classique sera utilisé afin de comparer les performances réalisées par le produit développé et c'est ainsi dans ce cadre que j'ai effectué mes travaux lors de ce stage.

1.2.2 Objectifs et sujets principaux

Le stage a comporté deux aspects principaux :

1. un travail de recherche théorique,
2. un travail opérationnel de test ou d'application de la théorie introduite.

Voici quelques exemples de sujets sur lesquels j'ai travaillé durant ce stage : les algorithmes de gradient boosting, **les algorithmes d'optimisation d'hyperparamètres**, la synthétisation de données pour la simulation de problématiques réelles, la détection de biais dans les modèles de machine learning. Ce mémoire portera essentiellement sur les algorithmes d'optimisation d'hyperparamètres.

La partie opérationnelle c'est déroulée essentiellement par l'intermédiaire de l'outil Python et de ces modules statistiques.