CHALLENGE DATA :
APPRENDRE LES SCIENCES DES DONNÉES PAR LA PRATIQUE

Gilles Wainrib[1] & Stéphane Mallat[2]

[1] *Département d'Informatique, Ecole Normale Supérieure, 45 rue d'Ulm 75005 Paris, gilles.wainrib@ens.fr*

[2] *Département d'Informatique, Ecole Normale Supérieure, 45 rue d'Ulm 75005 Paris, stephane.mallat@ens.fr*

Résumé. « Challenge DATA » est une initiative soutenue par la Fondation des Sciences Mathématiques de Paris et par l'Ecole Normale Supérieure. Conçue pour créer un pont entre les formations académiques en data science, le monde de la recherche et les applications réelles et innovantes, la plateforme « Challenge DATA » accueille des compétitions de machine learning, avec des projets proposés par des start-ups françaises, des entreprises industrielles ainsi que des laboratoires de recherche. Ces projets s'adressent aux étudiants effectuant leur formation dans des master de machine learning et data science, mais aussi aux data scientists et aux chercheurs intéressés. Chaque challenge est un problème supervisé de type classification ou régression, organisé sous forme d'une compétition à la manière de Kaggle. La plateforme a accueilli cette année 12 challenges couvrant un large éventail de types de données, de domaines d'applications et de problématiques de machine learning.

Mots-clés. Compétitions de machine learning, open innovation

Abstract. « Challenge DATA » is an initiative supported by the Fondation des Sciences Mathématiques de Paris and by Ecole Normale Supérieure, to bridge the gap between real applications, research and teaching in data sciences. The Challenge DATA platform hosts machine learning challenges, based on data provided by start-ups, innovative companies, medical centers and scientific experiments. They are addressed to students, data scientists and researchers in machine learning and statistics. Each challenge is a supervised classification or regression problem, organized as a competition according to the Kaggle procedure. This year 12 challenges covering a wide range of data, applications and machine learning problems have been hosted by the platform.

Keywords. Machine learning competitions, open innovation

1 Introduction

Avec le support de la Fondation des Sciences Mathématiques de Paris, l'École Normale Supérieure (Ulm) a organisé durant l'année 2015-2016 des compétitions de machine learning, destinés aux étudiants de Master et doctorat de la région Parisienne.

Le but de la plateforme challengedata.ens.fr est d'offrir à des enseignements liés à la statistique, au traitement de données, ou à l'apprentissage, des projets algorithmiques et numériques, avec des données réelles, mises à disposition par des start-ups, des entreprises et des laboratoires de recherche. Les problèmes proposés sont de type classification, régression ou prédiction, et couvrent un spectre très large d'applications: images, sons, capteurs, données médicales, financières, atmosphériques, etc.

Les principaux masters de la région parisienne sont partenaires de ce challenge, dont les Master Math Vision Apprentissage (MVA) de l'Ecole Normale Supérieure (Cachan), Data Science de l'Ecole Polytechnique, MASH de l'université Paris-Dauphine et le Master de Mathématiques et Applications de Paris 6 avec son Certificat Big Data.

L'objectif pédagogique est d'exposer les élèves à des tâches réelles de traitement de données, qui sont pré-formatées pour simplifier la définition du problème. Ils peuvent ainsi comparer les différentes approches avec un concours entre élèves (et chercheurs), suivant la procédure adoptée par le site Kaggle[1]. Le site Web évaluera les performances des algorithmes des élèves sur des données de tests, en leur renvoyant un score et un classement. Les élèves intéressés pourront venir aux présentations des problèmes faites par les entreprises, mais ce n'est pas obligatoire car le site inclura toute l'information nécessaire pour les projets. Ces présentations pourront cependant permettre aux élèves de rentrer en contact avec des entreprises ou start-ups, en vue d'un stage potentiel.

Ce projet est l'extension d'une expérience pilote qui a très bien fonctionné dans le cadre du cours de S. Mallat dans le Master MVA (2014-2015). On a pu ainsi observer une forte demande à la fois du côté des élèves et des entreprises.

2 Challenge Data

2.1 Fonctionnement

Chaque créateur de challenge (entreprise, labo, start-up) propose un problème d'apprentissage supervisé (classification ou régression), en mettant à disposition pour les participants des données d'entraînement et des données de tests.

Le fonctionnement retenu pour la plateforme est similaire à celui introduit par Kaggle qui a connu un très grand succès ces dernières années. Les participants peuvent accéder aux données, séparées en entraînement/test, et soumettre des résultats sous forme de fichier .csv, dont la pertinence est quantifiée par un score de performance prédéfini pour chaque projet. Le classement des meilleurs étudiants permet ainsi d'identifier les meilleurs algorithmes propices à la résolution de chaque problème. Afin de limiter le sur-apprentissage, un nombre limité de soumissions par jour a été fixé à 2 et le score final de la compétition est calculé sur un dataset de validation qui n'a jamais été utilisé tout au long de la compétition.

[1] www.kaggle.com

Ces Challenges Data s'inscrivent dans un esprit d'échange scientifique, avec un partage de données et de résultats. Les données mises à disposition par les entreprises sont donc non-confidentielles. Les participants peuvent rédiger un rapport présentant leur approche algorithmique et leurs résultats, et ce rapport peut être mis à disposion des enseignants et des créateurs de challenge. Cependant, afin de préserver le travail des participants, ces derniers pourront conserver leur code et ne sont à aucun moment invités à le rendre public. Les projets peuvent être utilisés pour l'évaluation des élèves, suivant les modalités fixées par chaque cours (type de projet, contenu des rapports et des soutenances). Chaque enseignant peut créer son cours sur la plateforme et éventuellement spécifier une liste de projets pouvant être traités par les élèves dans le cadre du cours, et ainsi accéder directement aux rapports postés par les élèves.

Afin de créer un réseau dynamique réunissant enseignants, élèves et entreprises, les projets ont été présentés à l'École Normale Supérieure, entre octobre et décembre 2015. Cela a permis aux entreprises de rencontrer les étudiants, en vue d'un stage, et de discuter avec des chercheurs sur leurs problèmes techniques, et aux étudiants d'avoir un aperçu global des différentes problématiques rencontrées actuellement dans différents domaines d'applications.

2.2 Exemples de projets

Afin d'illustrer la diversité des projets proposés, le tableau 1 présente l'ensemble des projets de la saison 2015-2016.

Créateur du challenge	Titre du projet	Type de données	Type de problème
Dreem	Classification des états du sommeil	Série temporelle (EEG)	Classification
Sonoscanner	Détéction de bounding box en échographie obstétrique	Images	Regression
Cardiologs	Détection d'inversion d'electrodes	Série temporelle (ECG)	Classification
Shihab Lab	Classification d'EEG du cortex auditif	Série temporelle (EEG)	Classification
Oze Energies	Prédiction de la consommation énergétique	Série temporelle	Régression
Regaind	Prédiction de la qualité d'une photo	Images	Régression
Quantmetry / SNCF	Prédiction de pannes sur les trains SNCF	Données non structurées	Classification
Plume Labs	Prédiction de la pollution atmosphérique	Série temporelle	Régression
Capital Fund Management	Prediction du volume des transactions financières	Série temporelle	Régression
Université de Toulon	Projet Bird : classification de chants d'oiseau	Audio	Classification
Dassault Systèmes 3DS	Sensor reduction	Série temporelle	Regression
Reminiz	Reconnaissance de visages dans les films	Images	Classification

Tableau 1 : Les projets proposés lors de la session 2015-2016 du Challenge Data.

2.3 Résultats

Ces challenges offrent aux entreprises la possibilité, d'avoir accès aux résultats des algorithmes les plus performants pour tous les types de données. Lors de l'expérience pilote menée dans le cadre du master MVA en 2014-2015, les étudiants ont amélioré les résultats existants dans 50% des cas et plusieurs d'entre eux ont ainsi été recrutés en stage puis en CDI dans les start-ups concernées.

En 2015-2016, nous avons également observé que les meilleurs étudiants parvenaient dans la majorité des cas à approcher ou à dépasser les performances obtenues en interne par les créateurs des challenges.

3 Conclusions et perspectives

La première saison de Challenge Data a permis à environ 500 participants, plus de 30 enseignants et une douzaine de start-ups, entreprises et laboratoires de recherche de se rencontrer, d'apprendre et d'expérimenter autour de problèmes concrets d'analyse de données prédictive. Ainsi, les compétitions ont permis aux étudiants d'améliorer leurs compétences, aux créateurs de challenges de rencontrer des collaborateurs potentiels et d'accéder aux dernières innovations algorithmiques.

Afin d'améliorer la plateforme pour les années à venir, plusieurs pistes sont à l'étude. En particulier, il nous apparaît important d'améliorer la communication entre les participants, les créateurs de challenge et les enseignants, avec par exemple la mise en place d'un forum ou d'un système de collaboration de code informatique. Cependant, cela ouvre alors la question de l'équilibre à trouver entre compétition et collaboration dans un contexte d'enseignement académique.