Challenges d'analyse de données : une formation par la pratique transversale et multidisciplinaire en science des données

Jean-Baptiste Durand

CHALLENGES D'ANALYSE DE DONNÉES : UNE FORMATION PAR LA PRATIQUE TRANSVERSALE ET MULTIDISCIPLINAIRE EN SCIENCE DES DONNÉES

Jean-Baptiste Durand
Laboratoire Jean Kuntzmann, Inria Mistis, Université Grenoble Alpes, Grenoble, France
Jean-Baptiste.Durand@imag.fr

Résumé. Nous présentons ici un projet pédagogique axé sur des challenges d'analyse de données, qui sera mis en œuvre à l'Université Grenoble Alpes à la rentrée 2017-2018. Ces challenges, transversaux à plusieurs formations, se veulent multidisciplinaires. Nous présentons l'organisation de ces challenges, leur positionnement dans le tissu universitaire grenoblois, et les projets de plateforme et de salle multimodale sur lesquelles ils s'appuient.

Mots-clés. Challenge, compétition, analyse des données, science des données, apprentissage statistique.

Abstract. A teaching project base on data challenges is presented in this communication. This project will begin in school year 2017-2018 at Université Grenoble Alpes. The challenges will be interdisciplinary and cross-educational. The organization of the challenges is described, as well as their positioning within Grenoble University community, and the projects of platform and multimodal classroom on which they rely.

Keywords. Challenge, competition, data analysis, data science, machine learning.

1 Positionnement des challenges

La Comue UGA (Université Grenoble Alpes) fait face à une demande croissance de formations faisant intervenir la science des données à différents niveaux (IUT, master, doctorat et recherche), et dans des contextes disciplinaires variés (statistique, mathématiques appliquées, informatique, traitement du signal, physique, biologie ou autre). Cette demande est le fait de la généralisation de l'utilisation d'internet et de l'essor des objets connectés. Par conséquent, les acteurs économiques et sociaux aussi bien que les particuliers sont confrontés au stockage et à l'analyse de quantités massives de données, susceptibles de contenir des informations pertinentes quant à divers phénomènes d'intérêt [1]. Les disciplines de la fouille de données et de l'apprentissage statistique reposent sur le pari que l'analyse automatique de telles quantités massives de données permettra de comprendre des connections et des corrélations entre de tels phénomènes, même s'ils peuvent sembler éloignés, et aidera à la prise de décision. Cependant, la transformation de masses de données en masses de connaissances à forte valeur ajoutée (par rapport à des objectifs de valorisation scientifique, économique, industrielle, et notamment en termes de prise de décision) n'est possible qu'à condition de développer des modèles qui fournissent des sorties interprétables, et que leurs utilisateurs aient les clés nécessaires à leur interprétation (autrement dit soient capables d'un couplage fin entre les propriétés mathématiques et algorithmiques des méthodes, et leur sémantique dans un contexte applicatif donné).

L'accroissement des performances des méthodes de fouille de données ces dernières années a mis en évidence leur intérêt, et a généré une forte augmentation de la demande de scientifiques des données dans les établissements publics et dans l'industrie. La formation d'un nombre adéquat d'étudiants avec des compétences adaptées à la fois en informatique, statistique, et avec des expériences de mise en œuvre dans des problèmes réels traitant des données de complexité significative est à présent un enjeu fort de l'Université. Eu égard à une relative défiance de certains étudiants vis-à-vis

soit des mathématiques, de la statistique ou de l'informatique, il s'agit de les aider à compléter leur formation dans ces domaines à la fois par la théorie et la pratique, afin de les amener à ces types de débouchés.

Répondre à une telle demande de formation est un défi pour l'équipe pédagogique de science des données à l'UGA, qui a souhaité proposer des solutions permettant à la fois d'impliquer les étudiants dans leur formation, et de mutualiser les moyens humains et matériels utilisés. Cette réponse est basée sur quatre axes, qui font l'objet d'un projet « Transformations pédagogiques et Plateformes "Learning-by-doing » et d'un projet Cross-Disciplinary Program (CDP) « DATA@UGA (Sciences des données) », financés par l'Idex « Université Grenoble Alpes : Université de l'innovation ». L'axe présenté ici est celui des challenges d'analyse de données. Cette approche est à même d'impliquer plus fortement les étudiants dans leurs études et de les plonger dans un contexte proche de celui de l'entreprise ou de la recherche.

Les challenges de données ont récemment connu un fort développement, qui s'est accompagné de la création de plateformes adaptées pour les héberger (kaggle.com, dextra.sg, drivendata.org entre autres). Pour les entreprises ou les instituts publics, ils sont une manière de sous-traiter un problème dont la résolution ne peut être menée en interne (l'exemple le plus connu étant la compétition Netflix de 2009 pour la prédiction de score d'appétence de consommateur pour la location de DVD). Pour les compétiteurs, la participation est un moyen de mettre en valeur leurs compétences auprès de recruteurs, une activité ludique… ou une opportunité de faire leurs premières armes en analyse de données. Ce dernier aspect a conduit les universités à créer leurs propres sites de challenges (challengedata.ens.fr), ou les sites existants à proposer une version éducative (inclass.kaggle.com)[2].

Cependant l'utilisation de ces challenges, popularisées dans d'autres formations et à l'international [2,3], reste assez marginale au niveau licence / master dans la communauté grenobloise. Nous souhaitons à la fois en faire la promotion et en faciliter l'accès aux étudiants et aux enseignants.

2 Modalités prévisionnelles de mise en œuvre

Les challenges d'analyse de données participent à un principe d'enseignements professionnalisants et innovants, qui s'inscrivent dans une pédagogie par projet/problème. Les challenges seront des compétitions basées sur des données réelles, souvent complexes, provenant de situations concrètes. Ils permettront aux étudiants de mettre en œuvre, par groupe, des méthodes de fouille de données et d'apprentissage automatique, mais aussi des compétences nécessaires dans le monde professionnel : communication, travail en groupe, compréhension d'un problème issu d'un domaine de spécialité étranger aux étudiants.

Afin de favoriser ces aspects, les équipes d'étudiants et d'enseignants constituées pour les projets seront multidisciplinaires, et feront appel à des compétences complémentaires en termes de modèles mathématiques, méthodes d'analyse et d'optimisation, systèmes d'information, clusters de calcul, sans oublier l'apport de problématiques réelles issues de divers domaines nécessitant l'intervention d'experts (notamment en santé, biologie, météorologie, imagerie, sciences humaines et sociales, langues). Une cellule d'enseignants permettra la coordination des interventions, favorisant ainsi les mutualisations et le fait d'impacter de nombreux étudiants. L'un des enjeux est donc de réunir des spécialistes qui puissent échanger et contribuer à faire vivre cette initiative pédagogique, ainsi que les enseignements et infrastructures associés, de sorte qu'elle puisse être modulable et utilisable dans différents contextes applicatifs et pédagogiques.

Le principe même des challenges visera également à rendre les étudiants plus acteurs de leurs apprentissages. Notre projet vise notamment à favoriser l'auto-évaluation sur des points spécifiques de la démarche du challenge d'analyse de données, au-delà de l'évaluation globale fournie par les classements des participants au challenge. Des cours spécifiques et séances de tutorat encadrées seront proposés afin de guider les étudiants vers des méthodes d'apprentissage automatique et de fouille de données qui paraissent incontournables dans le monde de la recherche et de l'entreprise,

et de les former aux langages de programmation disponibles sur les infrastructures proposées.

Cette pédagogie sera mise en œuvre à la rentrée 2017, en investissant de manière prioritaire dans les filières où la science des données a une place prépondérante : masters de statistique, mathématiques appliquées, informatique et traitement du signal, ainsi que le département STID (de 130 à 140 étudiants visés sur l'ensemble de ces formations). La promotion actuelle d'étudiants en IUT STID a déjà participé à un challenge de visualisation de données en 2017, et sera la première composante à bénéficier de la cellule après septembre 2017.

Afin de faciliter l'accès des enseignants et des étudiants aux challenges, notre projet comprend une plateforme et une salle multimodale. En effet, les plateformes existantes de challenges d'analyse de données sont contraignantes quant à la nature des problèmes proposés, et se cantonnent à la prédiction de données quantitatives (régression) ou qualitatives (classification) à partir de prédicteurs, dans un cadre supervisé (la plateforme fournit des données d'apprentissage pour calibrer le modèle). Les plateformes ne fournissent ni d'outils de programmation, ni de puissance de calcul, mais ne servent qu'à héberger les données, à classer les participants et à les aider à interagir via des forums.

Des plateformes existantes offrent une palette de fonctionnalités dont la nôtre pourra s'inspirer (notamment celle mise en place à l'ENS Paris, https://challengedata.ens.fr/, voir aussi [3]). Cependant, ouvrir les challenges d'analyse de données à une palette plus vaste d'étudiants et de fournisseurs de données tout en facilitant leur travail, nécessite de repenser les fonctionnalités de notre future plateforme. Elle devra en effet permettre l'inscription des étudiants et enseignants, le travail collaboratif en termes d'échange d'information et de développement de code, le dépôt de problèmes d'analyse de données avec des structures (séquences, graphes, …) et des questions variées (identification de relations de dépendances ou de causalité entre diverses quantités, …). Afin que les étudiants disposent d'une base de comparaison équitable, l'exécution du code sera réalisée sur cette même infrastructure, qui devra donc être adaptée aux grandes masses de données.

La salle multimodale comprendra des îlots de tables et PC à demeure, organisés autour de prises réseaux et d'alimentation pour des ordinateurs portables. Les îlots pourront être séparés par des cloisons mobiles, et chaque mur disposera d'un écran interactif, ou bien d'un écran et un tableau mobile. Ces équipements permettront aussi bien la projection de mini-cours ou tutoriels pour tous les groupes à la fois, que la communication au sein des groupes à l'aide de graphes ou de formules.

En termes pédagogiques et organisationnels, l'enjeu est de mutualiser des séances de formation méthodologique de manière transverse aux différentes formations impliquées, de faire reconnaître les challenges d'analyse de données pour la validation d'ECTS, des séances pratiques encadrées et non-encadrées en salle multimodale avec des créneaux communs, et également de mutualiser la salle en question.

Dans une démarche d'amélioration continue, la plateforme, les formations et les challenges développés seront testés sur des publics variés (math-info, biologie, psychologie, économie). Ces tests seront réalisés de façon séquentielle. Le groupe de travail des enseignants-chercheurs, qui se réunira régulièrement, analysera les différents retours pour continuer à enrichir et améliorer la diversité des challenges en termes à la fois applicatifs et méthodologiques (types de structures de données et de problèmes à résoudre). Une page à destination des encadrants gardera l'historique des jeux de données utilisés dans les challenges, avec leur caractéristiques, les meilleurs scores, les codes et résultats numériques associés, afin de constituer une base de données de challenges que nous pourrons à terme rendre accessible à des extérieurs au projet. Les retours viseront également à améliorer notre plateforme en termes de qualité et quantité de fonctionnalités offertes : forums, développement et partage de code, nombre de nœuds du cluster de calcul par exemple.

Le projet sera évalué par l'Idex sur la base d'indicateurs : essentiellement le nombre d'étudiants et de formations différentes impactées, et l'utilisation de la plate-forme.

3 Intégration dans le tissu grenoblois et perspectives d'évolution

Ce projet de challenges d'analyse de données s'appuie sur les trois autres volets du projet Idex formation « Data@UGA » :

I. Communication, promotion à l'international et cartographie des formations

II. Mutualisation et création de ressources pédagogiques dédiées à la science des données

III. Diffusion de la culture statistique, essaimage et formation des doctorants, enseignants et chercheurs.

I. Les masters de mathématiques appliquées, d'informatique et de traitement du signal concernés par les challenges sont internationaux et ce projet bénéficiera donc de la démarche mise en œuvre dans le volet I de « Data@UGA »
II. Les étudiants des masters concernés par les challenges, notamment les étrangers, ont une formation assez variée à leur arrivée à la COMUE UGA. Un enjeu est donc d'arriver à leur fournir un socle de connaissances qui les rende aptes à contribuer aux challenges. Nous envisageons des remises à niveau en auto-formation, en s'appuyant sur le volet II du projet « Data@UGA ».

III. Les formations et infrastructures seront également accessibles aux doctorants – sous forme d'inscription en certificat universitaire - et aux chercheurs, qui participent déjà à des challenges organisés par DataInstitute@grenoble. Ils auront également accès à des séminaires, écoles, autres challenges internationaux, et aux camps d'entraînement intensifs (« Boot camps ») pluridisciplinaires de l'institut. En effet, des compétitions sont organisées chaque année, notamment dans le cadre de conférences nationales ou internationales – c'est le cas notamment de la conférence sur l'apprentissage organisée cette année en 2017 à Grenoble, où un challenge autour de l'extraction des entités nommées a eu lieu de janvier à juin, en collaboration avec Viseo. Des échanges de données sont prévus entre la plateforme de challenges et la plateforme de diffusion de données PerSCiDO (http://persyval-platform.imag.fr) du labex Persyval. Ainsi, ce projet offre de nouvelles opportunités d'interactions entre chercheurs de la COMUE UGA et constitue pour les étudiants un tremplin vers le monde de la recherche.

Ainsi, les challenges prétendent contribuer à la visibilité nationale et internationale de l'Idex et de la COMUE UGA, notamment via l'ouverture à terme de cette plateforme à des acteurs extérieurs à ces entités.

Suite à la formation par les challenges d'étudiants de masters de statistique, mathématiques appliquées, informatique, traitement du signal, et STID, nous mèneront une politique d'essaimage de cette action pédagogique. Via la mise en place de projets multidisciplinaires issus du monde socio-économique, nous toucherons à la fois les étudiants des parcours de master statistique, de chargé d'études économiques et statistiques et les étudiants de Sciences Po. Ces projets allieront des spécialistes de communication, de management, d'analyse de données, d'économétrie et d'informatique (projet COMET-LabCampus[1]).

Un autre projet concerne les étudiants de M2 de l'École supérieure du professorat et de l'éducation (ESPE) et du master de statistique. Les étudiants de M2 de l'ESPE sont amenés à mettre en place une expérimentation pédagogique, à recueillir des données pendant leur stage et à les analyser. Chaque groupe d'étudiants de l'ESPE sera binômé avec un étudiant en statistique pour l'aide à la mise en place de l'expérience et de l'analyse.

1 http://www.communaute-univ-grenoble-alpes.fr/fr/formation/comet-labcampus-709559.htm

Bibliographie

[1] Cointot, J.-C. et Eychenne, Y. (2014) *La révolution Big data*. Dunod, Paris.
[2] Molla, D. (2013) Overview of the 2013 ALTA Shared Task. In : *Proceedings of the Australasian Language Technology Association Workshop*, 132–136. S. Karimi & K. Verspoor, Eds. 4–6 December 2013, Queensland University of Technology, Brisbane, Australia.
[3] Todeschini, A. et Genuer, R. (2015) Compétitions d'apprentissage automatique avec le package R rchallenge. In : *47èmes Journées de Statistique de la SFdS*, 1–5 Juin 2015, Lille, France.