Wikistat 2.0: Ressources pédagogiques pour l'Intelligence Artificielle

Philippe Besse, Brendan Guillouet, Béatrice Laurent

Wikistat 2.0 : Ressources pédagogiques pour la Science des Données et l'Intelligence Artificielle

Philippe Besse[1], Brendan Guillouet[2] et Béatrice Laurent[3]

TITLE

Wikistat 2.0 : Educational Resources for Artificial Intelligence

RÉSUMÉ

Big data, science des données, *deep learning*, intelligence artificielle, sont les mots clefs de battages médiatiques intenses en lien avec un marché de l'emploi en pleine évolution qui impose d'adapter les contenus de nos formations professionnelles universitaires. Quelle intelligence artificielle est principalement concernée par les offres d'emplois ? Quelles sont les méthodologies et technologies qu'il faut privilégier dans la formation ? Quels objectifs, outils et ressources pédagogiques est-il nécessaire de mettre en place pour répondre à ces besoins pressants ? Nous répondons à ces questions en décrivant les contenus et ressources opérationnels dans la spécialité Mathématiques appliquées, majeure Science des Données, de l'INSA de Toulouse. L'accent est mis sur une formation en Mathématiques (Optimisation, Probabilités, Statistique) fondamentale ou de base associée à la mise en œuvre pratique des algorithmes d'apprentissage statistique les plus performants, avec les technologies les plus adaptées et sur des exemples réels. Compte tenu de la très grande volatilité des technologies, il est impératif de former les étudiants à l'autoformation qui sera leur outil de veille technologique une fois en poste ; c'est la raison de la structuration du site pédagogique `github/wikistat` en un ensemble de tutoriels. Enfin, pour motiver la pratique approfondie de ces tutoriels, un jeu sérieux est organisée chaque année sous la forme d'un concours de prévision entre étudiants de masters de Mathématique Appliquées pour l'IA.

Mots-clés : *science des données, intelligence artificielle, apprentissage statistique, données massives, enseignement, jeux sérieux.*

[1]Université de Toulouse, INSA ; Institut de Mathématiques de Toulouse, UMR CNRS 5219 ; philippe.besse@math.univ-toulouse.fr

[2]Université de Toulouse INSA ; IRT Saint Exupéry

[3]Université de Toulouse, INSA ; Institut de Mathématiques de Toulouse, UMR CNRS 5219 ; béatrice.laurent@math.univ-toulouse.fr

ABSTRACT

Big data, data science, deep learning, artificial intelligence are the key words of intense hype related with a job market in full evolution, that impose to adapt the contents of our university professional trainings. Which artificial intelligence is mostly concerned by the job offers ? Which methodologies and technologies should be favored in the training pprograms ? Which objectives, tools and educational resources do we needed to put in place to meet these pressing needs ? We answer these questions in describing the contents and operational ressources in the Data Science orientation of the speciality Applied Mathematics at INSA Toulouse. We focus on basic mathematics training (Optimization, Probability, Statistics), associated with the practical implementation of the most performing statistical learning algorithms, with the most appropriate technologies and on real examples. Considering the huge volatility of the technologies, it is imperative to train students in seft-training, this will be their technological watch tool when they will be in professional activity. This explains the structuring of the educational site `github/wikistat` into a set of tutorials. Finally, to motivate the thorough practice of these tutorials, a serious game is organized each year in the form of a prediction contest between students of Master degrees in Applied Mathematics for IA.

Keywords : *data Science, artificial intelligence, statistical learning, big data, teaching, serious game.*

1 Introduction

1.1 Battage médiatique, scientifique et commercial

Big Data Analytics, Data Science, Machine Learning, Deep Learning, Intelligence Artificielle, un battage médiatique (*buzz word*) en chasse un autre, reflets ou écumes de disruptions technologiques importantes et surtout continuelles. Quels objectifs et choix pédagogiques engager pour anticiper les contenus de nos formations et les modes d'acquisition des compétences afin de préparer efficacement l'intégration des nouveaux diplômés ? C'est à ces questions que nous tâchons d'apporter des éléments de réponses, non pas des réponses théoriques ou des déclarations d'intention, mais plutôt des retours d'expériences et de réalisations en constante (r)évolution au sein de la spécialité Mathématiques Appliquées de l'INSA de Toulouse.

Il est évidemment important de communiquer avec les bons intitulés et les étudiants de l'INSA l'ont bien compris en se déclarant *data scientist* sur leur CV depuis 2013. Mais les bons choix d'investissement, ceux pédagogiques qui prennent du temps et engagent sur la durée, ne peuvent être pris en s'attachant à l'écume des mots, même inlassablement soulevée par des rapports officiels, média ou *hashtags* des réseaux sociaux. Il ne suffit pas de changer un intitulé de cours ou de diplôme.

Alors que les ressources pédagogiques, les MOOCs, SPOCs, tutoriels, se déversent à profusion sur internet, que devient le rôle d'un enseignant et plus précisément d'un enseignant / chercheur ? Certes contribuer à produire de la connaissance par la recherche mais, en responsabilité pédagogique, une fonction essentielle consiste à prioriser des choix. Sous l'écume médiatique, quelles sont les méthodes, les technologies, les algorithmes, dont les performances, donc la diffusion, motivent le temps et l'implication nécessaires à leur intégration dans un cursus académique inexorablement contraint par le volume horaire ?

La pression médiatique n'est pas seule en jeu, il faut noter aussi celle, académique, de publication : *publish or perish*, qui conduit à la production de milliers d'articles décrivant l"invention" de centaines de méthodes, algorithmes, librairies et de leurs très nombreuses variantes incrémentales, alors qu'en pratique, il faut reconnaître que les différences de performance n'apparaissent pas toujours significatives. Lire à ce sujet les articles de Hand (2006) et Donoho (2015) très critiques envers les algorithmes d'apprentissage récents, et pas seulement pour opacité et manque d'interprétabilité.

Notons aussi la pression commerciale ou publicitaire des entreprises, *start-up* ou grands groupes, disputant les parts d'un marché en forte croissance mais très volatil ou versatile. Chaque année depuis 2012 et motivés par des besoins de visibilité économique, Matt Turck et ses collaborateurs de la société *Firstmark* proposent une représentation graphique du paysage ou de l'écosystème, devenu fort complexe, des entreprises traitant de données massives ou plutôt maintenant de données massives et d'intelligence artificielle. La figure 1 résume celui de 2018. Ils tâchent chaque année de prendre en compte les créations, disparitions, fusions, des entreprises du domaine.

1.2 Quelle Intelligence artificielle ?

Les entreprises ayant appris à stocker, gérer massivement leurs données depuis 10 ans, la phase suivante concerne leur analyse pour leur valorisation et l'aide à la décision, voire de la décision automatique. Après s'être appelée "*big data analytics*" puis "*data science*" cette phase fait maintenant référence à une pratique d'*intelligence artificielle* (IA), appellation largement médiatisée, notamment depuis les succès remarquables en reconnaissance d'images depuis 2012, en traduction automatique, d'AlphaGo en 2016 ou autour des expérimentations de véhicules autonomes.

L'IA n'est pas une invention récente car cette discipline ou plutôt cet ensemble de théories et techniques est apparue conjointement avec le développement des tous premiers ordinateurs (ENIAC en 1943), eux-mêmes conséquences des efforts, durant la deuxième guerre mondiale, pour produire rapidement des abaques de balistique puis réaliser les calculs de faisabilité de la première bombe atomique. L'objectif initial était la simulation des comportements du cerveau humain. C'est aussi en 1943 que Mc Culloch (neurophysiologiste) et Pitts (logicien) ont proposé les premières notions de *neurone formel*. Notons le début de la théorisation de l'IA avec les travaux pionniers d'Alan Turing en 1950 et la première apparition de la notion de *perceptron*, le premier réseau de neurones formels, par Rosenblatt en 1957. Manque de moyens de calcul et d'algorithmes pertinents, l'approche connexionniste de l'IA est mise en veilleuse durant les années 70 au profit de la logique formelle (*e.g.* calcul des prédicats du premier ordre) comme outil de simulation du raisonnement. Les *systèmes experts* associant base de connaissance (règles logiques), base de faits et moteur d'inférence ont connu un certain succès, notamment avec le développement du langage *Prolog*, mais on butté sur la complexité algorithmique explosive des problèmes NP complets. Ce fut alors, au début des années 80, le retour massif de l'approche connexionniste avec le développement de l'algorithme de rétropropagation du gradient qui a ouvert la possibilité, en lien avec des moyens de calculs suffisamment performants, de l'apprentissage de réseaux de neurones complexes. Le développement de l'IA s'est ensuite focalisé