L'Ensai a 20 ans. A quelles compétences former les ingénieurs de demain?

Ronan Le Saout[1]
[1] ENSAI-CREST, Campus de Ker-Lann, Rue Blaise Pascal, 35170 BRUZ. ronan.le-saout@ensai.fr

Résumé. L'Ensai, école d'ingénieurs alliant compétences théoriques et opérationnelles en modélisation statistique, a célébré ses 20 ans en 2016. Ces 20 dernières années ont été marquées par l'émergence du concept de « Data Science » et une concurrence accrue avec d'autres acteurs institutionnels. L'Ensai entame donc une réflexion sur les métiers et compétences auxquels ses ingénieurs se doivent d'être formés, tout en conservant un enseignement de la statistique en tant que science sous-domaine des mathématiques appliquées. Cette communication synthétise les éléments collectés et présente les évolutions en cours à l'Ensai. Cette démarche peut paraître proche de celle engagée à l'INSA Toulouse il y a quelques années (Besse et Laurent 2014). Elle se distingue néanmoins par la place qu'occupent les compétences associées d'informatique et d'économie à l'Ensai.

Mots-clés. Data Science, enseignement, machine learning

Abstract. Ensai, an engineering school combining theoretical and operational skills in statistical modeling, celebrated these 20 years in 2016. The last 20 years have been marked by the emergence of the concept of "Data Science" and increased competition with other institutional actors. The Ensai therefore begins a reflection on the professions and skills to which its engineers must be trained, while maintaining a teaching of statistics as a science under the field of applied mathematics. This paper summarizes the elements collected and presents some insights concerning the evolution of the program. This approach may seem similar to that initiated at INSA Toulouse a few years ago (Besse and Laurent 2014). It should be nevertheless distinguished thanks to the associated skills of computer science and economics taught at Ensai.

Keywords. Data Science, teaching, machine learning

1 Introduction

L'Ensai a célébré en 2016 les 20 ans de son implantation à Rennes, sur le Campus de Ker-Lann. Plusieurs événements y furent associés, conférences sur le machine learning et le métier de statisticien mais aussi un Data Challenge en partenariat avec la Sncf. De manière concomitante, une démarche de réflexion sur l'évolution des enseignements et de la formation des ingénieurs s'est engagée, s'appuyant sur une enquête auprès des anciens élèves, l'avis des entreprises partenaires, d'experts professionnels et académiques, ainsi que par l'étude des offres d'emploi proposées sur Internet. L'émergence du concept de « Data Science » et la concurrence accrue avec d'autres acteurs institutionnels obligent en effet à une réflexion sur les métiers et compétences auxquels les ingénieurs de l'Ensai se doivent d'être formés, tout en conservant un enseignement de la statistique en tant que science sous-domaine des mathématiques appliquées. Par ailleurs, l'Ensai forme également des statisticiens publics (fonctionnaires de l'INSEE), l'INSEE a engagé en parallèle une réflexion sur la formation de ces cadres (Erkel-Rousse, Joly, Tassi 2017).

L'Ensai est une école d'ingénieurs alliant compétences théoriques et opérationnelles en modélisation statistique. En dernière année, six filières de spécialisation sont proposées, l'une à orientation statistique, une autre à orientation informatique, quatre autres plus thématiques (économie de la santé, marketing, gestion des risques, bio-statistique). L'originalité de sa formation réside historiquement dans des compétences associées fortes en informatique (avec une introduction de la programmation orientée objet dès le début du cursus et un projet de développement informatique en Java) et en économie (modélisation micro et macro-économique, et étude de la particularité des données économiques). Environ 100 ingénieurs sortent diplômés chaque année.

Cette communication synthétise les éléments collectés et présente les évolutions en cours à l'Ensai. Cette démarche peut paraître proche de celle engagée à l'INSA Toulouse il y a quelques années (Besse et Laurent 2014). De ce point de vue, certains choix d'évolution des enseignements renforçant l'apprentissage statistique sont proches. La démarche se distingue néanmoins par la place qu'occupent les compétences associées d'informatique et d'économie à l'Ensai. La discussion se concentre donc sur 3 points : les compétences en informatique à associer à une formation en statistique, les compétences en économie à associer à une formation en statistique et le rôle des nouvelles formes de projets du type Data Challenge lorsque des projets classiques existent déjà dans le cursus.

2 La démarche engagée à l'Ensai

Plusieurs éléments ont été collectés: enquête auprès des anciens élèves, avis des entreprises partenaires, d'experts professionnels et académiques, étude des offres d'emploi proposées sur Internet. Ces éléments ont appuyé de premières évolutions de la 1ère année du cursus. Pour les 2 années correspondant aux années de Master à l'Université, la réflexion est encore en cours. Le cœur des compétences reste le socle scientifique statistique, avec une alliance des compétences théoriques et opérationnelles. Ces réaménagements se traduisent par un meilleur cadencement des compétences, des compétences en informatique plus directement reliées aux compétences en statistique, un renforcement des enseignements en apprentissage statistique, du caractère opérationnel de la formation à travers la mise en œuvre de projets, et des « soft skills » (communication écrite et orale, ouverture vers d'autres domaines de compétences tels que le droit).

Qu'est-ce qu'un Data Scientist ? Analyse des offres d'emploi

Pour décrire son métier, un « Data Scientist »[1] a proposé une définition par la pratique sur sa page LinkedIn. Il a « scrappé » (i.e. téléchargé depuis le Web) les informations contenues sur les pages LinkedIn comprenant le terme Data Scientist. Il a ensuite listé les 10 principales compétences associées : Data Mining, Machine learning, R, Python, Data Analysis, Statistics, SQL, Java, matlab, Algorithms. Il a enfin tracé un graphe relationnel des associations entre ces compétences faisant apparaître des clusters de compétences, soulignant trois sphères : la modélisation statistique à partir d'une approche mathématique et théorique des données (compétences en machine learning, data mining, data analysis et statistics), la création, l'exploration, et l'implémentation de modèles (compétences en R, Python, Matlab), l'informatique (compétences en Java, C++, Algorithmes et Hadoop). Il est rare d'être un expert dans ces trois catégories qui apparaissent dans l'environnement professionnel d'un Data Scientist. Un cursus de formation en Data Science doit dispenser des compétences dans ces trois sphères, mais l'ambition en termes de compétences à acquérir pour l'ensemble des élèves (i.e. le tronc commun) doit être réfléchie.

Pour l'ENSAI, Myriam Vimond (enseignante-chercheuse en statistique) a effectué un travail complémentaire en analysant **des** offres d'emploi proposées à des statisticiens (Data Analyst, Data Scientist) sur internet de juillet 2015 à novembre 2015 en distinguant les profils de statisticien et d'informaticien. Certains langages (R, python) apparaissent comme communs. D'autres (programmation orientée objet (Java ou C++), Architectures distribuées (Hadoop, Spark, Hive, Pig, Impala, HBase)) sont maîtrisés par les informaticiens mais restent uniquement des compétences basiques pour les statisticiens. C'est l'inverse pour les compétences en machine learning, méthodologies et logiciels statistiques.

Avis des étudiants, des employeurs et d'experts

Des auditions d'experts académiques et de l'entreprise ont souligné que l'Ensai devrait former dans l'idéal des ingénieurs capables de mener un projet de A à Z, de la collecte et du nettoyage des données à sa restitution sous forme d'applicatif en passant par la réflexion méthodologique. Développer une expertise sur l'ensemble du processus à l'issue d'un diplôme de niveau Master n'est pas possible. Il faut donc que l'ensemble des étudiants aient des bases sur la collecte et le nettoyage des données (y compris les sources hétérogènes, les API…), les applicatifs informatiques, voire des notions juridiques, de management, d'économie d'entreprise... Les diplômés doivent pouvoir être « agiles », dialoguer avec différentes sphères et s'auto-former. Le cœur des compétences doit rester le socle scientifique statistique, et donc la modélisation, mais en orientant les enseignements vers plus d'apprentissage statistique (et moins de statistique classique). Les compétences informatiques et statistiques doivent marcher ensemble. Il faut développer l'autonomie des étudiants, à travers les projets. La mise en place d'un data challenge dans le cursus (valorisé par des crédits ECTS) doit être réfléchie.

Concernant l'avis des employeurs, deux des conclusions sont la faible capacité des étudiants à communiquer auprès de non-spécialistes et l'utilisation désormais courante de Python, que ce soit pour usage informatique ou statistique. Les jeunes diplômés ont par ailleurs le sentiment de faiblement mobiliser les connaissances en économie.

1 Ferris Jumah 2014, LinkedIn

3 La place des compétences associées en informatique et en économie

L'objet de cette partie n'est pas de donner un avis conclusif mais de détailler les réflexions en cours à l'Ensai.

Concernant les compétences en informatique, l'Ensai avait historiquement des enseignements en Programmation orientée objet en Java tournés vers le développement et la production informatique. Il a été décidé de mieux lier les enseignements de statistique et d'informatique à travers un projet d'analyse et de traitement de données permettant de mettre en œuvre les compétences acquises au sein de la première année, notamment ceux de Bases de données et d'Algorithmique. Le nouveau projet se substitue à un projet plus classique de développement informatique. Il pourrait se traduire par:
1) l'importation et le nettoyage à l'aide d'un script Python ou d'une API de données publiques (portails opendata) ;
2) la mise en œuvre d'un modèle statistique simple sur ces données à l'aide de plusieurs algorithmes et de la définition d'objets simples (les méthodes pourront provenir des cours de statistique et d'optimisation) ;
3) le calcul de la complexité des solutions ;
4) un rendu sous forme visuelle (page Html, applicatif).

Concernant les compétences en économie, l'une des difficultés est le ressenti d'une faible mobilisation de ces compétences en début de carrière. Or l'économie vise à prendre du recul sur l'utilisation de données mesurant des comportements humains et à améliorer l'évaluation d'effets causaux. Une enquête auprès d'étudiants plus anciens va être effectuée pour mieux appréhender cette question.

Enfin, concernant les projets, des réserves ont vu jour sur les Hackatons et Data Challenges, et sur comment concilier un travail en temps court et la rigueur scientifique. Même avec l'adoption de métriques d'évaluation (erreurs de prévision...), il convient de distinguer les apports pédagogiques d'un Data Challenge et ceux des projets classiques.

Ces réflexions et ces changements visent à prendre en compte les évolutions du métier de statisticien, en ne le limitant pas à un rôle opérationnel. Le caractère scientifique des enseignements apparaît ainsi sur l'ensemble de la chaîne statistique, tant d'un point de vue appliqué que théorique, en écho aux évolutions du métier de statisticien (Donoho 2015).

Bibliographie

[1] Besse, P. et Laurent, B. (2000), *De statisticien à Data Scientist – Développements pédagogiques à l'Insa de Toulouse*, Statistique et Enseignement, 7(1), 75–93.
[2] Donoho D. (2015), *Fifty Years of Data Science*, Mimeo.
[3] Erkel-Rousse, H., Joly, P. et Tassi P. (2017), *Formation initiale des attachés*, Rapport de l'Inspection Générale de l'INSEE.
[4] Ferris J. (2014), *The Data Science Skills Network*, LinkedIn.