■ Introduction au Machine Learning,

à Mines ParisTech, le 28 mai 2019

L'école des Mines ParisTech nous a accueillis le 28 mai 2019 pour un stage LIESSE intitulé *Introduction au Machine Learning*, qui fait appel à la science des data et l'analyse convexe, à l'intersection entre les statistiques et l'informatique. Ce stage a été entièrement animé par l'enseignante-chercheuse Chloé-Agathe Azencott, devant une quinzaine d'entre nous. C'est une (petite) partie du cours qu'elle donne aux étudiants des Mines qui choisissent ce cursus optionnel, en passe de devenir une partie du tronc commun dans la réforme des enseignements que l'école est en train de terminer pour la rentrée 2019.

Elle nous a présenté avec bonne humeur et clarté six exposés sur l'apprentissage statistique : analyse en composantes principales, minimisation du risque empirique, régularisation, reseaux de neurones artificiels multicouches, méthodes à noyaux, arbres de décision et forêts aléatoires. Le premier exemple éclairant la démarche est celui de la recette de gâteau : au départ, au lieu d'avoir des ingrédients et une recette pour produire un gâteau, on part du principe qu'on a une grande quantité de gâteaux et d'ingrédients et qu'on essaie de retrouver la recette.

Parmi les objectifs, elle a donné un certain nombre d'exemples comme la reconnaissance de chiffres manuscrits ou la prédiction de la quantité de clics générés par un article publié, ainsi que faire reconnaître à un ordinateur une image de chat ou non, un spam ou non, une image de cellule cancéreuse ou non, c'est-à-dire des problèmes balayant un large spectre.

Au départ, nous disposons d'un grand nombre de données X à exploiter et d'étiquettes y (ou classification), l'objectif étant de retrouver la fonction f qui à X associe y. Les questions présentées étaient : comment gérer la très grande dimension des données traitées (clustering, réduction de la dimension), ce qui nous a menés à un problème de valeurs propres, comment minimiser le risque empirique, avec plusieurs modèles de régression, comment mesurer la performance d'un algorithme pour généraliser ce qu'il a appris, et comment éviter à l'algorithme de sur-apprendre (on peut trouver une fonction très bien adaptée aux données présentes mais très mauvaise ailleurs, d'où la nécessité d'entraîner l'algorithme sur une partie des données, de le tester sur une autre partie et de le valider sur une troisième). Nous avons également vu des modèles de réseaux de neurones artificiels, de machines à vecteurs de support (SVM) et d'arbres de décision. Elle nous a également présenté les outils pour s'essayer au Machine Learning, dont le module `scikit-learn` de Python, développé par l'Inria.

Nous avons été très bien accueillis par l'école, dont le directeur adjoint et le directeur des études, Matthieu Mazière, sont venus déjeuner avec nous et répondre à nos questions variées sur l'école et ses formations. Nous les remercions très chaleureusement.

Aliénor Burel