

<https://archined.ined.fr>

**Eric Biernat, Michel Lutz, 2017, Data science :
fondamentaux et études de cas, Machine
learning avec Python et R, Paris, Eyrolles,
296 p.**

Elisabeth Morand

Version

Libre accès

POUR CITER CETTE VERSION / TO CITE THIS VERSION

Elisabeth Morand, 2018, "Eric Biernat, Michel Lutz, 2017, Data science : fondamentaux et études de cas, Machine learning avec Python et R, Paris, Eyrolles, 296 p.", Population (édition française) 73: 404-405.

Disponible sur / Available at:

<http://hdl.handle.net/20.500.12204/AWgYFm58XMQCvuZmm6Zl>

BIERNAT Eric, LUTZ Michel, 2017, *Data science : fondamentaux et études de cas, Machine learning avec Python et R*, Paris, Eyrolles, 296 p.

Cet ouvrage est destiné à tous ceux qui veulent découvrir la *data science*. Comme l'indique le sous-titre de l'ouvrage, *Machine learning avec Python et R*, il est ici bien question d'une introduction au *machine learning*⁽¹⁾ par la pratique. L'ouvrage est composé de trois parties, la dernière étant exclusivement consacrée à l'étude de cas pratiques. L'ouvrage, très didactique, expose les aspects pratiques des algorithmes de *machine learning*, sans oublier de poser au préalable les éléments fondamentaux dans les deux premières parties.

Il se présente sous forme de fiches, ce qui permet une lecture très souple de différents concepts. Les premières fiches définissent le vocabulaire, ce qui peut avoir un intérêt même pour un public averti. Par exemple, la fiche sur la régression logistique permet d'avoir sur cette méthode un point de vue différent de ceux développés couramment dans les ouvrages de statistique ou de méthodes pour les sciences sociales.

L'organisation générale de l'ouvrage correspond à un apprentissage des méthodes mais peut aussi être utile dans un emploi quotidien, comme support aux analyses courantes. Les trois grandes parties dont il est composé correspondent chronologiquement aux trois grandes étapes que l'on doit suivre dans une analyse : définition du problème, choix de la méthode, enfin mise en œuvre.

Dans la première partie, très courte, les auteurs tiennent à préciser le type de problèmes auxquels tente de répondre l'ouvrage, ainsi que les outils nécessaires pour le faire. Il y a donc au sein de cette première partie un chapitre sur les logiciels qui sert aussi à définir le périmètre couvert par l'ouvrage.

La deuxième partie, scindée en 14 chapitres, présente les méthodes d'analyse du-*machine learning*. La progression y est très classique pour un ouvrage introductif. Cette partie présente séparément les analyses classiques et les analyses plus contemporaines. Parmi ces dernières, on trouvera les forêts aléatoires, le *gradient boosting* et le *support vector machine* (svm), qui sont souvent objet d'interrogation quand on débute dans le domaine. Pour ceux qui veulent, par exemple, comprendre les forêts aléatoires, la fiche est limpide et tient en seulement une dizaine de pages, sans omettre pourtant l'essentiel.

La troisième partie est la partie clé, bien qu'elle ne puisse se passer de la deuxième, ni surtout de la première partie. Elle décrit en détail de nombreux cas pratiques que les auteurs ont choisi de prendre sur la plateforme web Kaggle, ce qui permet au lecteur d'avoir un accès aux données, voire aux programmes. Au-delà des méthodes présentées dans la partie précédente, on trouvera ici des méthodes pratiques, par exemple une présentation de l'*online-learning* (p. 219), méthode utile si votre jeu de données est suffisamment important.

(1) L'apprentissage automatique ou apprentissage statistique, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyens algorithmiques plus classiques.

Les auteurs pointent aussi les forces et les faiblesses des deux composantes dominantes de la *data science*, la statistique et l'informatique, et estiment que la première aide à mieux appréhender les développements de la seconde. Ils pensent aussi, à juste titre, que les méthodes de visualisation devraient faire l'objet d'une plus grande diffusion. Certes, on ne peut qu'adhérer à cette idée, mais on aurait souhaité qu'elle s'accompagne d'éléments précis sur l'aide et la complémentarité que ces méthodes peuvent apporter. On regrettera aussi le manque de recours à un langage scientifique français. En effet, l'usage des termes anglo-saxons est omniprésent alors que des équivalents français existent.

Enfin, cet ouvrage offre une large bibliographie particulièrement adaptée et composée d'ouvrages et d'articles récents, mais aussi d'ouvrages classiques fondamentaux. Ajoutons qu'elle est d'une consultation facile car fournie au fur et à mesure qu'on avance dans la lecture des fiches. En définitive, ce manuel est une boîte à outils de référence pour débiter dans le *machine learning*, tout en étant agréable à lire. En effet, les auteurs ont une très longue pratique professionnelle des méthodes présentées ainsi qu'une longue expérience pédagogique, dont le lecteur bénéficie.

Elisabeth MORAND