Union des Professeurs de classes préparatoires Scientifiques

Introduction au machine learning avec Python

Le 23 octobre 2019

ENSTA ParisTech - Palaiseau

Effectif : 14

Auteur du CR : Valentin Raban

/ Présentation

Nous étions entre douze et quatorze enseignants de mathématiques et de physique-chimie à ces deux journées d'introduction au machine learning, les 23 et 24 octobre. Il s'agit de la deuxième édition de cette formation, l'ENSTA l'ayant déjà proposée l'année dernière.

Les deux journées sont indépendantes. La première a été encadrée par Zacharie Ales, qui a présenté une technique de réduction de dimension la matinée, et quelques méthodes de classification non supervisée l'après-midi. Chaque demi-journée consistait en une heure et demi de présentation par l'intervenant puis une heure et demi de pratique avec la librairie **scikit-learn** de Python. La deuxième journée fait l'objet d'un rapport séparé.

/ Déroulé

Sur l'exemple d'un jeu de données répertoriant des individus décrits chacun par de nombreuses caractéristiques (âge, taille, poids, ...), on identifie rapidement les difficultés liées à la masse d'information : visualiser les données (les graphiques sont au maximum en 3D), les interpréter (de plus en plus délicat au fur et à mesure que le nombre de caractéristiques croît) et les généraliser (plus de variables tend à augmenter le risque de sur-apprentissage). Partant de ce constat, une approche simpliste consiste à éliminer directement certaines des caractéristiques. On peut néanmoins être tenté de réduire la taille des données plus intelligemment. C'est l'idée des méthodes projectives, dont l'analyse en composantes principales (ACP) fait partie. Il s'agit d'effectuer un changement de base en choisissant les nouveaux axes de telle sorte à ce que le maximum d'information soit représenté par un minimum d'axes, appelés alors principaux. On peut apprécier la qualité de cette représentation par la notion de part d'inertie, liée à la variance des données sur les axes principaux. Un point négatif de cette méthode est la difficulté d'interpréter la signification de ces nouveaux axes. Pour remédier à ce problème, le cercle des corrélations est un outil permettant de visualiser l'impact des caractéristiques primaires sur chacun des axes principaux (dans le cas où ils sont au nombre de deux). M. Ales a fourni de nombreux détails mathématiques et pragmatiques sur chacune de ces notions, puis les participants ont pu les mettre en pratique en suivant un énoncé explicite et bien étalé en difficulté. Les exemples ont notamment porté sur la classification de fleurs.

Après un déjeuner offert par l'ENSTA au restaurant de Polytechnique avec M. Ales, l'après-midi a repris par une présentation de méthodes d'apprentissage non supervisé sur des problèmes de classification. Parmi celles-ci, on distingue les méthodes hiérarchiques et les méthodes de partitionnement. Une méthode simple de classification hiérarchique d'un ensemble consiste à regrouper itérativement les deux éléments les plus proches. Une diversité d'algorithmes naît alors spontanément de la définition de « plus proches éléments ». Des algorithmes plus sophistiqués ont aussi été présentés (ROCK et Chameleon). M. Ales a ensuite discuté deux méthodes de partitionnement : l'algorithme ***k-means*** et celui de propagation d'affinité. Leurs points forts et leurs faiblesses ont été explicitement illustrés. La mise en pratique qui a suivi est revenue sur la classification de fleurs.

Tout le long de la journée, M. Ales s'est montré très disponible et les discussions ont été riches. Les participants, pour la plupart néophytes dans le domaine, ont apprécié une introduction au machine learning qui a balayé plusieurs aspects de la discipline sans jamais être superficielle. Les supports de

cours et de TP contiennent une bibliographie pour ceux souhaitant aller plus loin dans les notions présentées. Certains concepts, par exemple les méthodes hiérarchiques de classification non supervisée, paraissent adaptés pour une présentation en deuxième année d'IPT.

Valentin Raban