Union des Professeurs de classes préparatoires Scientifiques

INTRODUCTION AU MACHINE LEARNING AVEC PYTHON

Le 24 octobre 2019

ENSTA ParisTech - Palaiseau

Auteur du CR : Jean-Baptiste Flament

Présentation

Ce bilan fait suite à celui écrit par Valentin Raban concernant la journée du 23 octobre. La seconde journée s'est déroulée sous l'encadrement de David Filliat, directeur de l'Unité Informatique et Ingénierie des Systèmes de l'ENSTA, et centrée cette fois sur l'apprentissage supervisé le matin, et sur les réseaux de neurones l'après-midi. La structure était la même que la veille (1h30 cours + 1h30 TP), avec à nouveau la bibliothèque scikit-learn le matin, mais avec PyTorch l'après-midi, utilisée sur la plateforme Google Colaboratory permettant des calculs sur cartes graphiques (GPU).

Déroulé

L'apprentissage supervisé a pour but de construire une approximation d'une fonction ayant beaucoup de paramètres (pouvant aller jusqu'à plusieurs millions de variables), et dont l'ensemble image peut être un ensemble continu (problème de régression), un ensemble discret (problème de classification) ou binaire (problème de détection). Cette construction s'appuie sur des bases de données déjà traitées, chères en pratique et obtenues auprès de sociétés spécialisées, en évaluant la qualité de l'approximation par comparaison de ses résultats à ceux contenus dans la base de données. L'apprentissage se décompose de plus en une partie d'entrainement et une partie de validation, afin d'éviter une trop grande spécificité de l'approximation à l'échantillon de données utilisé, et garantir sa généralité.

Après une présentation théorique des concepts de l'apprentissage supervisé, plusieurs algorithmes d'entrainement sont présentés (Support Vector Margin -SVM, séparateur à vaste marge en français-, arbres de décision et forêts d'arbres aléatoires) ainsi que les outils adaptés à leur évaluation (*e.g. :* détection : précision/rappel, classification : matrices de confusion, arbres : entropie, ...), et M. Filliat insiste à chaque fois très clairement sur les avantages et inconvénients de chaque méthode et sur les problèmes auxquelles elle est adaptée. La partie pratique, assez bien guidée dans un premier temps, permet de se familiariser avec les fonctions déjà implémentées dans scikit-learn et d'observer les algorithmes et outils présentés plus tôt sur des exemples de classifications de fleurs et de reconnaissance de chiffres. Il était ensuite possible d'aller plus loin en essayant de traiter d'autres exemples de bases de données étant inclus dans la bibliothèque.

La journée reprend l'après-midi avec une présentation générale du deep learning et des réseaux de neurones, avant une présentation plus spécifique à la reconnaissance d'images. Datant des années 50, le concept de neurone artificiel et de réseaux de neurones s'est énormément développé depuis les années 2000, grâce notamment à l'augmentation de la puissance de calcul parallèle sur GPU. La présentation théorique se penche à nouveau sur les méthodes et outils utilisés (neurones, fonctions d'activation, perceptron multi-couches, poids, méthode de descente du gradient, inertie, cross-entropy...), avant de voir leur application à la reconnaissance images (problème de classification) et aux méthodes spécifiques à cette problématique, assez techniques, ayant pour but de réduire le nombre d'opérations nécessaires. Cette seconde partie donnant un contexte plus appliqué, elle a apporté un éclairage nouveau sur les concepts précédents, et mené à un long échange de questions avec M. Filliat. La partie pratique visait alors à classifier les images d'une base de données en s'appuyant sur des fonctions prédéfinies de PyTorch.nn. Elle permettait notamment d'appréhender les enjeux du déroulement de la procédure d'entrainement (seule intervention humaine dans cet apprentissage) et des différentes techniques la composant.

En conclusion, cette introduction tout à fait accessible au machine learning permet d'avoir un bref aperçu de ce vaste domaine, tout en en présentant les enjeux et les principales difficultés conceptuelles. L'aspect informatique

lui-même est tout à fait abordable, une grande partie des fonctions étant implémentées dans scikit-learn et PyTorch.