Intelligence artificielle pour l'astrophysique à l'époque du big-data

Marc HUERTAS-COMPANY
Mdc - LERMA

**Syllabus**:

L'astronomie entre définitivement dans l'ère du big-data. Suite à cette transition, la façon dont on extrait et interprète l'information des données évolue rapidement et notamment l'utilisation de l'intelligence artificielle apparait comme incontournable. Ceci est d'autant plus vrai que le domaine de l'apprentissage automatique vit une époque dorée avec l'émergence de l'apprentissage profond (deep-learning).

L'objectif principal de la formation est de présenter les principes du machine learning (apprentissages supervisé, non supervisé, réseaux de neurones etc..), en partant de l'apprentissage dit classique pour arriver principalement à l'apprentissage profond et son utilisation sur de grands volumes de données. Une importance particulière sera donnée à l'illustration par des exemples concrets afin de comprende comment ces techniques peuvent être employées en astrophysique.

Le module comporte deux parties. Une partie théorique où des intervenants de différentes disciplines présentent les principes de l'apprentissage automatique. Une deuxième partie pratique où les étudiants seront confrontés à un vrai *data challenge* sur un cas pratique réel pour lequel ils devront trouver une solution utilisant l'intelligence artificielle (e.g. classification de galaxies, recherche de signatures spectrales, exo-planètes, recherche de lentilles gravitationnelles, détection d'objets etc).

A la fin de la formation de 5 jours, l'étudiant sera familiarisé avec les bases du machine learning (notamment de l'apprentissage profond) et sera capable d'appliquer un réseau simple à un ensemble général de données. Le langage de programmation utilisé est python.

**Liste d'intervenants:**

Liste encore à définir.

L'idée étant d'avoir des intervenants inter-disciplinaires provenant de plusieurs instituts.

1. Astrophysiciens utilisateurs des techniques illustrant leur utilisation dans leur discipline.

2. Spécialistes de machine learning et traitement d'images notamment de l'Ecole des Mines et du Centre for Data Science à Saclay permettant d'expliquer les concepts techniques de base et le fonctionnement.

**Programme:**

Le module durera 4 à 5 jours et comportera des cours et une mise en pratique par des exemples. Notamment l'idée est d'organiser des "compétitions" du style organisées dans la communauté de machine learning.

La forme définitive doit encore être finalisée avec les différents intervenants. Les cours seront impartis en anglais.

A titre d'exemple, voici une organisation possible:

DAY 1 (~6hrs):

General presentation: What is machine learning? What is big-data and why they go hand by hand?
Types of learning Supervised vs. unsupervised
Classical machine learning algorithms - Classification, Regression…

DAY 2 (~6hrs):

Neural networks, deep-learning
Evaluation of models
Identification of problem, overfitting etc.

DAYS 3/4 (~12 ++ hrs):

Data challenge. The students will be confronted to a real example presented in form of a data challenge. Following a standard approach they will be given a dataset to train and a blind set to test. They will have 2 days to come up with a solution for the proposed problem. All solutions will be uploaded to a common platform that will compare, evaluate and ranks the results.

DAY 5 (~4hrs):

Oral presentations by students of the proposed solutions. Common discussion on differences and ways to combine different approaches.

**Logistique**

La contrainte principale est d'avoir accès à une salle de TP avec ordinateurs et accès internet. Pour la mise en place du challenge nous utiliserons des plateformes existantes permettant centraliser et comparer les résultats. Notamment des collègues du Center for Data Science ont développé des solutions versatiles permettant ce type de fonctionnement.

**Pré-requis**

Bonne connaissance de Python

**Justification:**

L'astronomie n'est pas une exception, et comme beaucoup d'autres disciplines, entre définitivement dans l'ère du big-data. C'est le cas pour la plupart des spécialités. En l'espace de 15 ans, nous sommes passés d'avoir à notre disposition quelques données isolées pour un nombre très réduit d'objets, à des mesures couvrant une grande partie du spectre électromagnétique pour des millions, voire des billions de galaxies grace notamment aux sondages profonds. L'avenir s'avère encore plus brillant avec les projets de la prochaine décennie tels que LSST ou EUCLID, pour n'en citer que deux, sur lesquels travailleront un partie importante des nouveaux docteurs.

Une chose s'avère inévitable: il va falloir modifier la façon dont on extrait et interprète l'information des données et notamment l'utilisation de l'intelligence artificielle apparait comme incontournable. Tout simplement parce que des méthodes plus *manuelles* s'avèrent impossibles. Ceci d'autant plus que l'intelligence artificielle vit une révolution. L'intérêt des grands géants technologiques cherchant à trouver des moyens efficaces de traiter l'avalanche d'information circulant dans le web, favorise le

développement rapide de nouvelles techniques de plus en plus efficaces. L'apprentissage profond (deep-learning) en est la preuve.