Apprentissage statistique
Projet : *Higgs Boson Machine Learning Challenge*

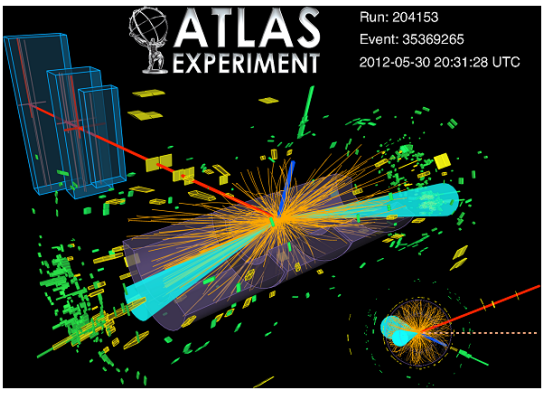
Olivier Schwander <olivier.schwander@lip6.fr>

À rendre pour le mercredi 16 janvier 2018 à 23h59 (heure de Paris)

On s'intéresse à la détection du boson de Higgs dans des données simulées de façon à reproduire le comportement de l'expérience ATLAS. Il s'agit d'un problème de classification binaire, ou de détection d'événement. Les deux classes sont :

— *background*
— *tau tau decay of a Higgs boson*

Il s'agit du projet Kaggle *Higgs Boson Machine Learning Challenge* https://www.kaggle.com/c/higgs-boson/. Comme indiqué sur la page du projet : **Aucune connaissance en physique des particules n'est nécessaire**. De plus, il n'est pas exigé que vous soumettiez vos résultats sur le site (mais ça n'est bien sûr pas interdit).



Consignes

Binômes

Ce projet doit être réalisé en binôme (sauf exception justifiée). Merci de m'indiquer vos choix de binômes avant les vacances.

Rendu final

Les documents suivants devront être rendu à la fin du projet :

— un **rapport** (au plus 10 pages, au format pdf),
— le **code** (les fichiers sources, dans une archive tar.gz ou zip).

Le rapport devra contenir une description complète de la démarche, depuis le chargement des données, leur analyse préliminaire, leur pré-traitement, l'application de méthodes, la sélection des paramètres

jusqu'à l'analyse et l'interprétation des résultats. Dans le cas de résultats négatifs, des commentaires pertinents sont bienvenus. Il est indispensable de donner une brève description de toutes les méthodes utilisées.

Le code pourra être réalisé dans le langage de votre choix (l'utilisation de Python, Numpy, Scipy, Pandas, Sklearn ou Keras est bien sûr conseillée). Une **documentation** (même succincte) devra être fournie pour expliquer comment faire tourner les différentes méthodes proposées et produire un fichier contenant les résultats (ce point est *particulièrement* important si le langage utilisé n'est pas Python, mais une documentation claire est indispensable dans tous les cas).

Une **soutenance** sera organisée fin janvier. Il vous sera demandé de présenter brièvement quelques points marquants de votre travail. L'exposé sera suivi de quelques questions.

Notation

Les points suivants seront pris en compte :

— la pertinence de la démarche,
— la qualité du rapport,
— la qualité du code,
— la qualité de l'exposé et des réponses aux questions.

La notation portera très peu sur les scores de classification obtenus mais beaucoup plus sur la pertinence de la démarche ainsi que sur la compréhension des méthodes utilisées. Vous êtes encouragés à utiliser toutes les méthodes qui vous sembleront utiles (qu'elles fassent ou non partie du cours, et qu'elles fassent ou non partie d'une bibliothèque existante) mais il est essentiel que chacune des méthodes utilisées soit comprise et puisse être expliquée (au moins dans ses grandes lignes) lors de la soutenance.

Remarques

Il n'est pas exigés que vous implantiez vous-même les méthodes utilisées, l'utilisation de bibliothèques est acceptable et encouragée.

Vous êtes encouragés à faire de la bibliographie autour du jeu de données pour trouver des idées de démarches et mieux comprendre les données.

Remarque importante : vous êtes très vivement encouragés à poser des questions, soit pendant les séances, soit par courriel.

Données

Les données sont disponibles sur la page du cours ainsi que sur la page http://opendata.cern.ch/record/328. Le fichier `atlas-higgs-challenge-2014-v2.csv.gz` contient les données étiquetées (818238 évènements, dont 279560 positifs). Chaque évènement est décrit par 30 features.

Les données, une description et d'autres informations utiles sont disponibles sur

— la page Kaggle https://www.kaggle.com/c/higgs-boson/data
— la page du Laboratoire de l'Accélérateur Linéaire (LAL) https://higgsml.lal.in2p3.fr/
— la page du CERN http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014

Un guide de démarrage est notamment proposé sur la page https://higgsml.lal.in2p3.fr/software/starting-kit/. Il est fortement recommandé de l'utiliser comme baseline.

Important Pour vos essais, vous aurez à définir vous même un protocole d'évaluation et une séparation entre données d'apprentissage et données de test.

Pour soumettre sur Kaggle, il faudra utiliser le fichier `training.zip` pour l'apprentissage et le fichier `test.zip` (pas étiqueté) pour le test.

Démarche (suggestions)

On donne ici quelques suggestions générales pour guider l'étude. Toutes les remarques ne s'appliqueront pas forcément à ce jeu de données précis et à cette tâche précise.

Cette démarche n'est évidemment pas linéaire : il faudra souvent retourner en arrière pour modifier certaines étapes du processus.

Analyse des données

La première étape est d'étudier manuellement les données, par exemple en traçant des histogrammes du nombre d'événements en fonction de chaque colonne.

C'est également l'occasion de tester quelques classifieurs très simples. Souvent des idées très simples se comporteront mieux que des méthodes plus compliquées mais peu adaptées.

Prétraitements et construction des descripteurs

La deuxième étape est de choisir les colonnes que l'on va utiliser pour la prédiction, et comment on va les utiliser. Certaines sont évidemment inutiles, pour d'autres l'utilité mérite d'être étudiée. On peut vouloir appliquer des transformations sur les valeurs : prendre la somme de deux colonnes, ou le carré d'une autre, etc.

D'autres prétraitements peuvent être utiles : discrétiser ou seuiller des valeurs, transformer des valeurs catégoriques en valeurs numériques, etc.

Choix des méthodes

Une jungle de méthodes d'apprentissage est disponible, il faut commencer par utiliser les méthodes les plus simples possibles (principe du rasoir d'Ockham) puis raffiner si nécessaire.

Une méthode qui fait gagner un petit peu sur le score au prix d'un temps de calcul beaucoup plus long, et d'une complexité plus grande n'est pas forcément une bonne idée.

Sélection des paramètres

Beaucoup de méthodes dépendent d'hyper-paramètres. Il est acceptable dans un premier temps de choisir ces paramètres à la main, en fonction des résultats. Dans un second temps, il faut utiliser une méthode de sélection automatique comme la validation croisée.

Évaluation

L'évaluation de la qualité d'un modèle se fera sur une base de test (qui n'aura bien sûr pas été utilisée pour l'apprentissage). Ici, on calculera la précision de la prédiction.

Il faut analyser les résultats :

— le score est-il meilleur que d'autres méthodes ? moins bon ?
— le score sur la base de test est-il beaucoup moins bon que celui obtenu sur la base d'apprentissage ? (risque de sur-apprentissage)

— quels sont les temps de calcul ? quelle la quantité de mémoire utilisée ? Un éventuel sur-coût est-il compensé par un gain significatif sur le score ?
— la méthode est-elle simple à mettre en œuvre et à expliquer ? ou est-ce un modèle très compliqué, extrêmement spécifique au jeu de données, difficile à adapter à un problème un peu différent ?