REDS - TP2 : Baseline et méthode d'ensemble

Laura Nguyen et Keyvan Beroukhim

2 octobre 2019

Toujours à partir des données issues du *ATLAS Higgs Boson Machine Learning Challenge 2014*, l'objectif durant ce TP est d'établir une baseline pour la prédiction du type d'évènement : *s* pour *signal* et *b* pour *background*.

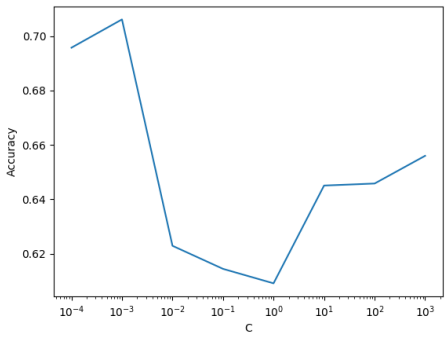
Avant de mettre en place des modèles d'apprentissage, les données doivent être nettoyées. On supprime d'abord les attributs `KaggleSet`, `Weight`, `KaggleWeight` et `EventId`. Ensuite, les features contenant plus de 40% de valeurs manquantes sont supprimées. Enfin, on retire les évènements à valeurs manquantes. On passe d'un dataset de 818238 à 693636 échantillons (environ 85% des données initiales) et de 35 à 20 attributs.

Par souci de temps de calcul, les expériences suivantes sont menées sur le dataset restreint à 10000 évènements. 70% de cette base est dédiée à l'entraînement et la validation des modèles, 30% au test.

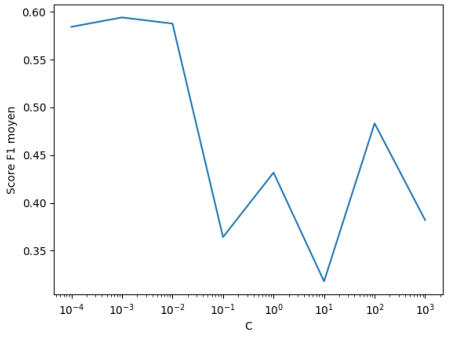
On évaluera les performances des modèles avec l'accuracy, le score F1 et le score AMS.

1 Baseline

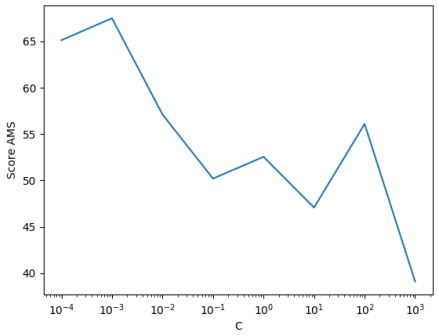
On établit dans un premier temps une baseline : les évènements seront classifiés à l'aide d'un SVM linéaire. Pour trouver la valeur optimale de l'hyperparamètre C, on effectue un *grid search* sur l'ensemble $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$. Pour chaque valeur, un SVM est entraîné par *cross-validation* sur 4 folds. Moyenner les scores obtenus sur chaque fold permet d'avoir une bonne estimation de la performance du modèle en apprentissage. Pour chacune des trois mesures d'évaluation, la figure 1 contient la courbe des performances moyennes produites avec chaque C.



(a) Accuracy



(b) Score F1



(c) AMS

Fig. 1: Performances moyennes en apprentissage en fonction de C

On récupère la valeur de C permettant d'obtenir le meilleur score F1 moyen sur les données d'entraînement. Un SVM est ensuite entraîné avec cette valeur sur l'ensemble d'entraînement. Pour chaque mesure d'évaluation, les performances en apprentissage et en test figurent dans le tableau 1.

	Accuracy	Score F1	AMS
Apprentissage	70.82%	58.47%	134.95
Test	70.69%	58.20%	87.74

Tab. 1: Scores obtenus en apprentissage et en test

2 Méthodes d'ensemble

On utilise maintenant plusieurs méthodes d'ensemble dans l'objectif d'améliorer les performances :

- *Bagging* avec un *Perceptron* comme classifieur faible
- *Random Forest*
- *AdaBoost*

Comme chacune de ces méthodes possède un nombre important d'hyperparamètres et qu'un grid search serait donc coûteux, on définit arbitrairement leurs valeurs.

Chaque classifieur est entraîné sur la totalité des données d'entraînement puis évalué sur l'ensemble de test. Les résultats en apprentissage et en test figurent dans le tableau 2.

	Accuracy		F1		AMS	
	Apprentissage	Test	Apprentissage	Test	Apprentissage	Test
Bagging	68.74%	67.17%	40.49%	38.44%	35.04	23.12
Random Forest	98.66%	78.07%	97.97%	69.10%	140.71	35.10
AdaBoost	80.43%	78.00%	74.19%	72.03%	58.06	35.06

Tab. 2: Performances obtenues par chaque méthode