TER-M1. Contributions à l'évaluation du HiggsML challenge

Contexte

Le challenge HiggsML[1] a formalisé une des composantes de l'analyse de l'expérience de physique des hautes énergies ATLAS comme un problème de classification avec des caractéristiques inhabituelles et difficiles[2]. La compétition a reçu un intérêt considérable de la communauté Apprentissage (~ 1800 participants, plus grand nombre de participation enregistré dans une compétition Kaggle sponsorisée).

L'objectif scientifique du TER est de participer à l'analyse immédiate des résultats du challenge, et à la mise en perspective à moyen terme des questions originales du challenge dans les concepts et mesures traditionnels en apprentissage

Travail demandé

Un premier travail (environ 40h) concerne l'analyse des résultats du challenge, dans le but de déterminer la validité statistique de la hiérarchie obtenue, en utilisant la méthodologie générale de (Dietterich[3]). On utilisera des tests statistiques classiques pour la comparaison par paires et en groupe. L'intérêt est dans la méthodologie de constitution de l'échantillon testé. Pour des raisons techniques, il devra être constitué par bootstrap. Le travail demandé est :

- une synthèse élémentaire sur l'approche boostrap pour le test à partir de (Davison & Hinkley[4]), destinée à des utilisateurs scientifiques non spécialistes ;
- le choix argumenté d'un logiciel de bootstrap (dans les librairies des environnements R, Matlab, Scikit learn, autres) ;
- un développement très léger, mais réutilisable et documenté interfacé avec la base de données de résultats du challenge, susceptible d'être exploité dans une plate-forme d'analyse de challenges ;
- les résultats concrets de l'analyse sur les résultats du challenge.

Le second travail (60h) concerne l'analyse de la complexité des données (sample complexity). Le but final est de fournir des métriques pertinentes sur la difficulté du problème d'apprentissage. A l'intérieur de ce très vaste domaine, le travail demandé est le suivant.

- Une synthèse élémentaire basée sur la lecture des passages non techniques de (Vayatis &Azencott[5]), (Bartlett et al.[6]), (Sabato et al.[7]) et le chapitre 3 de (Macia[8]).
- La mise en œuvre de la *Data Complexity Library*[5] sur les données d'entrée du challenge, sur une machine HPC. Les questions non triviales sont en particulier l'effet du traitement des données manquantes et surtout la complexité quadratique du calcul de certaines métriques. Au niveau du TER, on se limitera à évaluer la taille de données sur lesquelles une mesure est faisable en l'état, et à fournir les résultats. Le support technique pour l'accès à la machine sera fourni.

Encadrant

Cécile Germain, AO/TAO LRI/INRIA, co-organisateur du HiggsML Cjhallenge. cecile.germain@lri.fr

---

[1] https://www.kaggle.com/c/higgs-boson
[2] http://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf
[3] Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 7 (October 1998), 1895-1923.
[4] Davison, A. C.; Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
[5] Nicolas Vayatis and Robert Azencott. 1999. Distribution-Dependent Vapnik-Chervonenkis Bounds. In Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT '99), Paul Fischer and Hans-Ulrich Simon (Eds.). Springer-Verlag, London, UK, 230-240.
[6] Peter L. Bartlett , Olivier Bousquet , Shahar Mendelson. Local Rademacher complexities (2002). The Annals of Statistics 2005, Vol. 33, No. 4, 1497–1537
[7] Sivan Sabato , Nathan Srebro , Naftali Tishby. Tight Sample Complexity of Large-Margin Learning. NIPS 2010.
[8] Nuria Macia. Data complexity in supervised learning: A far-reaching implication. PhD Thesis , 2011.