https://lib.uliege.be https://matheo.uliege.be

Master thesis : Local permutation importances for random forests

Auteur : Weyders, Pierre-François
Promoteur(s) : Geurts, Pierre
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2020-2021
URI/URL : http://hdl.handle.net/2268.2/13302

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Master thesis : Local permutation importances for random forests

Pierre-François Weyders

August 2021

Interpreting the predictions of a model is a key component to the study of each model. This thesis focus on understanding each prediction of an ensemble of tree. The methods that are studied and compared provide an importance to each feature that summarizes the dependence of the model on the feature to correctly predict the outcome.

To the end of understanding the individual predictions of a model, the local methods defined by Saabas and SHAP are compared with two newly introduced methods based on the mean decrease in impurity and mean decrease in accuracy global methods defined by Breiman 2001 [1]. The methods are extensively studied on several datasets and the effect of the hyperparameters of the ensemble of trees on the rankings are studied too.

The rankings built by the local methods are compared to ground-truth rankings inferred from the weights of a linear regression problems. The performance of the methods strongly depends on the model's parameter. It is shown that the number of features considered at each split has a strong influence on MDA and MDI as these two methods extensively use the internal structure of a tree to infer local rankings. Over all tasks, we observed that SHAP and Saabas are well correlated. By studying the evolution of the rankings inferred by the local methods, we observed that MDI and MDA tend to converge to the same rankings as SHAP and Saabas as the number of feature considered at each split increases. In this context, methods provide a very similar interpretation on the prediction of the tree.

In the second part of the thesis, the methods are used to infer local regulatory networks and are shown to outperform previous state of the art methods. Expressions of genes are first generated using the dyngen[2] simulator. Current state of the art methods are presented: *pySCENIC, SSN* and *LIONESS+Pearson* and their results are compared with local methods. Random forests are trained on the expression of the genes and local interpretation methods are used to infer the regulatory activity that exists between the genes of each cells. The MDI method is shown to be very effective despite its computational simplicity.

References

[1] Leo Breiman. Importance measures derived from random forests. *PhD dissertation*, 6 2001.

[2] Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. dyngen: a multi-modal simulator for spearheading new single-cell omics analyses. *bioRxiv*, 2020.