The Journal of Open Source Software

RENT: A Python Package for Repeated Elastic Net Feature Selection

Anna Jenul[1], **Stefan Schrunner**[1], **Bao Ngoc Huynh**[2], **and Oliver Tomic**[1]

1 Department of Data Science, Norwegian University of Life Sciences **2** Department of Physics, Norwegian University of Life Sciences

DOI: 10.21105/joss.03323

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @maximtrp
- @arunmano121

Submitted: 29 April 2021
Published: 28 July 2021

Summary

Due to modern data acquisition techniques, the number of generated features in measurement data keeps increasing. This increase can make the analysis with standard machine learning methods difficult because of underdetermined systems where the dimensionality of the feature space (number of features) exceeds the dimensionality of the object space (number of observations). A concrete example of such a situation is data acquisition in the healthcare domain, where the number of patients (observations) suffering from a specific condition may be relatively low, but a lot of measurements (number of features) are generated for each patient to acquire a good understanding of the patient's health. A very common challenge is that not all features in a high dimensional space are equally important for predictive tasks — many might even be redundant. Feature selection deals with finding the most relevant features of a dataset. With help of appropriate methodology, feature selection can reduce (a) the complexity of and (b) noise in the dataset. More importantly, data interpretation of the model becomes easier with fewer features, which is of great importance within domains such as healthcare. Even though feature selection is a well-established research topic, relatively few approaches are focusing on the stability of the selection. The important question at hand is: can we trust that the selected features are really valid or is their selection very dependent on which observations are included in the data? Providing information on the stability of feature selection is vital, especially in wide data sets where the number of features can be many times higher than the number of observations. Here, the inclusion or exclusion of a few observations can have a high impact on which features may be selected.

Statement of Need

To get an understanding of which features are important and how stable the selection of each feature in the dataset is, a user-friendly software package is needed for this purpose. The RENT package, implementing the feature selection method of the same name (Jenul et al., 2021), provides this information through an easy-to-use interface. The package includes functionalities for binary classification and regression problems. RENT is based on an ensemble of elastic net regularized models, which are trained on randomly, iid subsets of the rows of the full training data. Along with selecting informative features, the method provides information on model performance, selection stability, as well as interpretability. Compared to established feature selection packages available in `R` and Python, such as `Rdimtools` (You, 2020) implementing Laplacian and Fisher scores or the scikit-learn feature selection module (Pedregosa et al., 2011) implementing recursive feature elimination and sequential feature selection, RENT creates a deeper understanding of the data by utilizing information acquired through the ensemble. This aspect is realized through tools for post hoc data analysis, visualization, and feature selection validation provided with the package, along with an efficient and user-friendly implementation of the main methodology.

Concept and Structure of RENT

At its core, RENT trains K independent elastic net regularized models on distinct subsets of the training dataset. Each subset is generated using the scikit-learn function `train_test_split()` which delivers an iid sample from the full training dataset. The sampling processes of different subsets are mutually independent, with the condition that a single data point can appear at most once in each subset. A data point, however, can appear in multiple subsets. The framework is demonstrated in Figure 1.

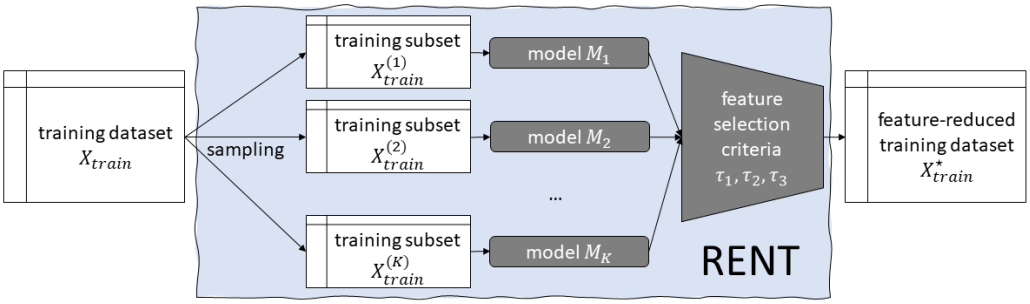

Figure 1: Summary of RENT method (Jenul et al., 2021).

Based on three statistical cutoff criteria τ_1, τ_2 and τ_3, relevant features are selected. While τ_1 counts how often each feature was selected over K models, τ_2 quantifies the stability of the feature weights — a feature where the K weight signs alternate between positive and negative is less stable than a feature where all weights are of a constant sign. The third criterion τ_3 deploys a Student's t-test to judge whether feature weights are significantly different from zero. The presented implementation builds on an abstract class `RENT_Base` with a general skeleton for feature selection and post hoc analysis. Two inherited classes, `RENT_Classification` and `RENT_Regression`, offer target-specific methods. The constructor of `RENT_Base` initializes the different user-specific parameters such as the dataset, elastic net regularization parameters, or the number of models K. After training, feature selection is conducted by use of the cutoff criteria. Deeper insights are provided by a matrix containing the cutoff criteria values of each feature, as well as a matrix comprising raw model weights of each feature throughout the K elementary model. For initial analysis of the results, the package delivers multiple plotting functions, such as a barplot of τ_1. Additionally, two validation studies are implemented: first, a model based on random feature selection is trained, while second, a model based on randomly permuted labels of the test dataset is obtained. Results of both validation models are compared to a model built with RENT features using Student's t-tests as well as empirical densities.

In addition to feature selection, RENT offers a detailed summary of prediction accuracies for the training objects. For each training object, this information can be visualized as histograms of class probabilities for classification problems or histograms of mean absolute errors for regression problems, respectively. For extended analysis, principal component analysis reveals properties of training objects and their relation to features selected by RENT. For computation and visualization of principal components, RENT uses functionality from the `hoggorm` and `hoggormplot` packages (Tomic et al., 2019).

Ongoing Research and Dissemination

The manuscript RENT - Repeated Elastic Net Technique for Feature Selection is currently under review. Further, the method and the package are used in different master thesis projects at the Norwegian University of Life Sciences, mainly in the field of healthcare data analysis.

Acknowledgements

We thank Runar Helin for proofreading the documentation.

References

Jenul, A., Schrunner, S., Liland, K. H., Indahl, U. G., Futsaether, C. M., & Tomic, O. (2021). *RENT – repeated elastic net technique for feature selection*. http://arxiv.org/abs/2009.12780

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Tomic, O., Graff, T., Liland, K. H., & Næs, T. (2019). Hoggorm: A python library for explorative multivariate statistics. *The Journal of Open Source Software*, *4*(39). https://doi.org/10.21105/joss.00980

You, K. (2020). *Rdimtools: Dimension reduction and estimation methods*. https://CRAN.R-project.org/package=Rdimtools