The Journal of Open Source Software

Confidence Intervals for Random Forests in Python

Kivan Polimis[1], **Ariel Rokem**[1], **and Bryna Hazelton**[1]

1 eScience Institute, University of Washington

DOI: 10.21105/joss.00124

Software

- Review
- Repository
- Archive

Summary

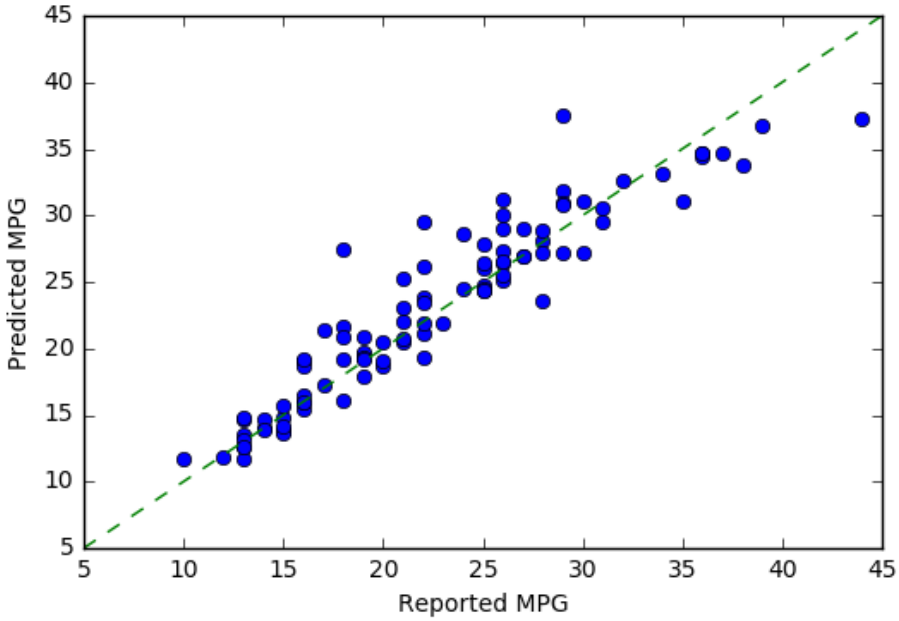
Random forests are a method for predicting numerous ensemble learning tasks. The variability in predictions is important for measuring and estimating standard errors and providing additional information about a metric's accuracy. `forest-confidence-interval` is a Python module for calculating variance and adding confidence intervals to `scikit-learn` random forest regression or classification objects. The core functions calculate an in-bag and error bars for random forest objects. Our software is designed for individuals using `scikit-learn` random forest objects that want to add estimates of uncertainty to random forest predictors. This module is an implementation of an algorithm developed by Wager, Hastie, and Efron (2014) and previously implemented in R (Wager 2016).

Examples gallery

The regression example uses a data-set from the UC Irvine Machine Learning Repository with features of different cars and their MPG (Lichman 2013). The classification example generates synthetic data to simulate a task like that of a spam filter: classifying items into one of two categories (e.g., spam/non-spam) based on a number of features. This module will work for matrices or `pandas` data frames. Then, `scikit-learn` functions split the example data into training and test data and generate a random forest object (regression or classifier). Our package's `random_forest_error` and `calc_inbag` functions use the random forest object (including training and test data) to create variance estimates that can be plotted (e.g. as confidence intervals or standard deviations). The inbag matrix that fit the data is set to `None` by default, and the inbag will be inferred from the forest. However, this only works for trees for which bootstrapping was set to `True`. That is, if sampling was done with replacement. Otherwise, users need to provide their own inbag matrix.

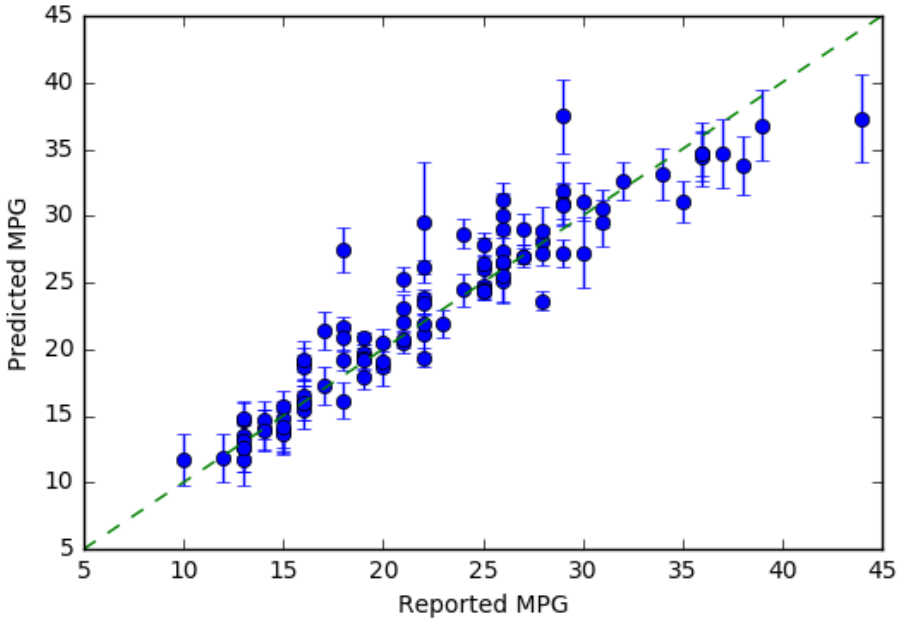
Regression example

No variance estimated



-

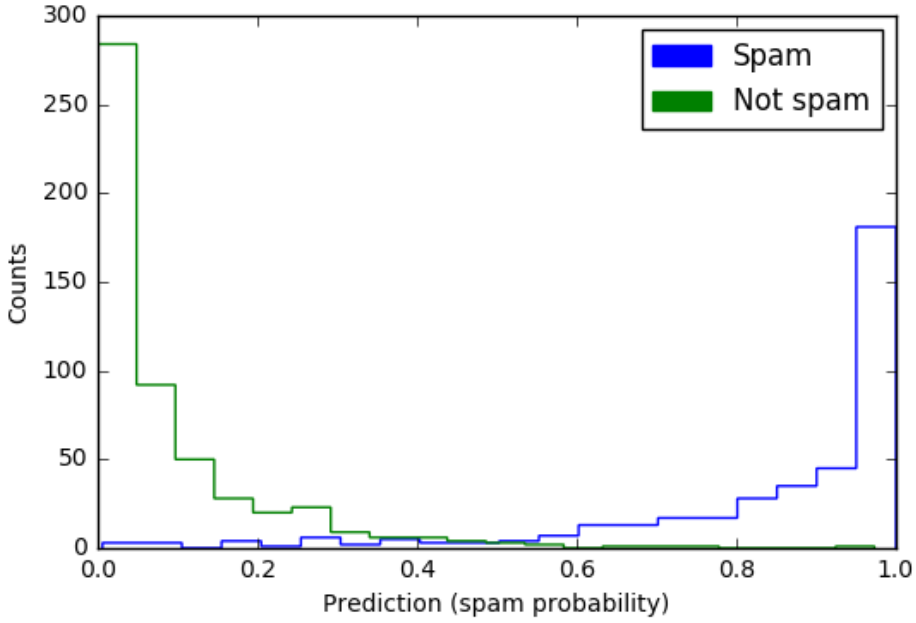
Plot with variance



-

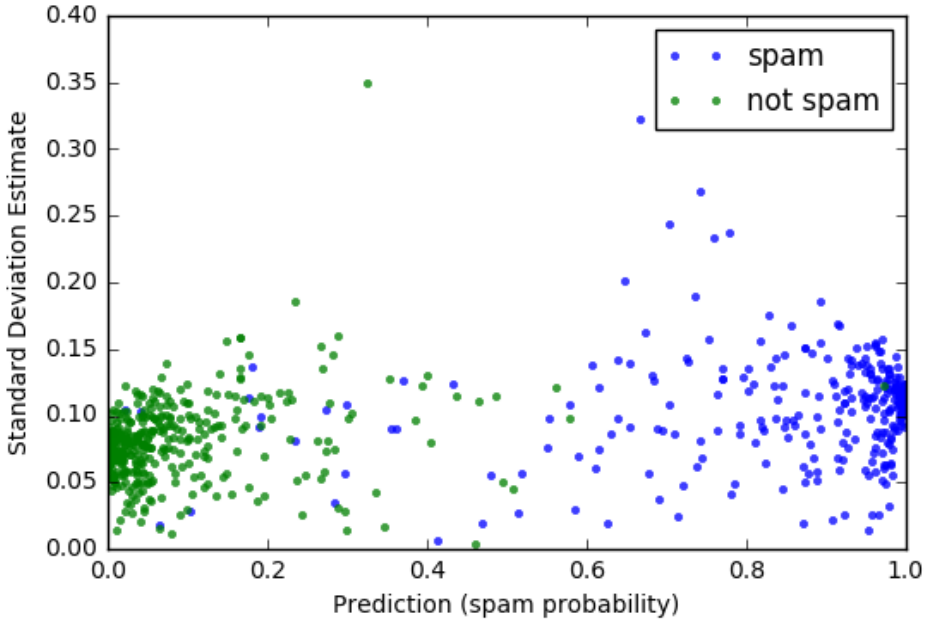
Classification example

No variance estimated



-

Plot with variance



-

Community guidelines

Contributions are very welcome, but we ask that contributors abide by the contributor covenant.

To report issues with the software, please post to the issue log Bug reports are also appreciated, please add them to the issue log after verifying that the issue does not already exist. Comments on existing issues are also welcome.

Please submit improvements as pull requests against the repo after verifying that the existing tests pass and any new code is well covered by unit tests. Please write code that complies with the Python style guide, PEP8

Please e-mail Ariel Rokem, Kivan Polimis, or Bryna Hazelton if you have any questions, suggestions or feedback.

References

Lichman, M. 2013. *UCI Machine Learning Repository.* University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Wager, Stefan. 2016. "randomForestCI." https://github.com/swager/randomForestCI.

Wager, Stefan, Trevor Hastie, and Bradley Efron. 2014. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife." *J. Mach. Learn. Res.* 15 (1): 1625–51. http://dl.acm.org/citation.cfm?id=2627435.2638587.