The Journal of Open Source Software

The vtreat R package: a statistically sound data processor for predictive modeling

John Mount[1] **and Nina Zumel**[1]

1 Win-Vector, LLC

DOI: 10.21105/joss.00584

Software

- Review
- Repository
- Archive

Submitted: 09 February 2018
Published: 24 February 2018

Summary

When applying statistical methods or applying machine learning techniques to real world data, there are common data issues that can cause modeling to fail. The `vtreat` package (Mount and Zumel (2018)) is an R data frame processor that prepares messy real world data for predictive modeling in a reproducible and statistically sound manner.

The package's objective is to produce clean data frames that preserve the original information, and are safe for model training and model application. `Vtreat` does this by collecting statistics from training data in order to produce a *treatment plan*. `Vtreat` then uses this treatment plan to process subsequent data frames prior to both model training and model application. The processed data frame is guaranteed to be purely numeric, with no missing or `NaN` values, and no string or categorical values. `Vtreat` serves as a powerful alternative to R's native `model.matrix` construct. The goals of the package differ from those of training harness systems such as `caret` (Jed Wing et al. (2017)) and unsupervised ad-hoc processing systems such as `recipes` (Kuhn and Wickham (2018)).

In particular vtreat emphasizes *safe but y-aware (supervised) pre-processing of data* for predictive modeling tasks. It automates:

- Treatment of missing values through safe replacement plus indicator column.
- Explicit coding of categorical variable levels as indicator variables.
- Robust handling of novel categorical levels (values seen during test or application, but not seen during training).
- Supervised re-coding of categorical variables with very large numbers of levels, using an approach similar to that described by J. Cohen and Cohen (1983).
- Cross validation to mitigate overfit and undesirable supervision bias.
- Optional significance-based and cross-validated variable selection.

`Vtreat` is careful to automate only domain-agnostic data cleaning steps that are to common to many applications. This intentionally leaves domain-specific processing to the researcher and their own appropriate tools.

The use of `vtreat` avoids the perils of ad-hoc data treatment, and provides a reproducible, documented, and citable data treatment procedure.

For more details and further discussion, please see our expository article Zumel and Mount (2017) and the package online documentation.

References

Cohen, Jacob, and Patricia Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum Associates, Inc.

Jed Wing, Max Kuhn. Contributions from, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2017. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Kuhn, Max, and Hadley Wickham. 2018. *Recipes: Preprocessing Tools to Create Design Matrices*. https://CRAN.R-project.org/package=recipes.

Mount, John, and Nina Zumel. 2018. *Vtreat: A Statistically Sound 'Data.frame' Processor/Conditioner*. https://doi.org/10.5281/zenodo.1173318.

Zumel, Nina, and John Mount. 2017. "Vtreat: A Data.frame Processor for Predictive Modeling." https://doi.org/10.5281/zenodo.1173314.