The Journal of Open Source Software

KLRfome - Kernel Logistic Regression on Focal Mean Embeddings

Matthew D. Harris[1]

1 AECOM Technologies

DOI: 10.21105/joss.00722

Software

- Review
- Repository
- Archive

Submitted: 14 April 2018
Published: 03 May 2018

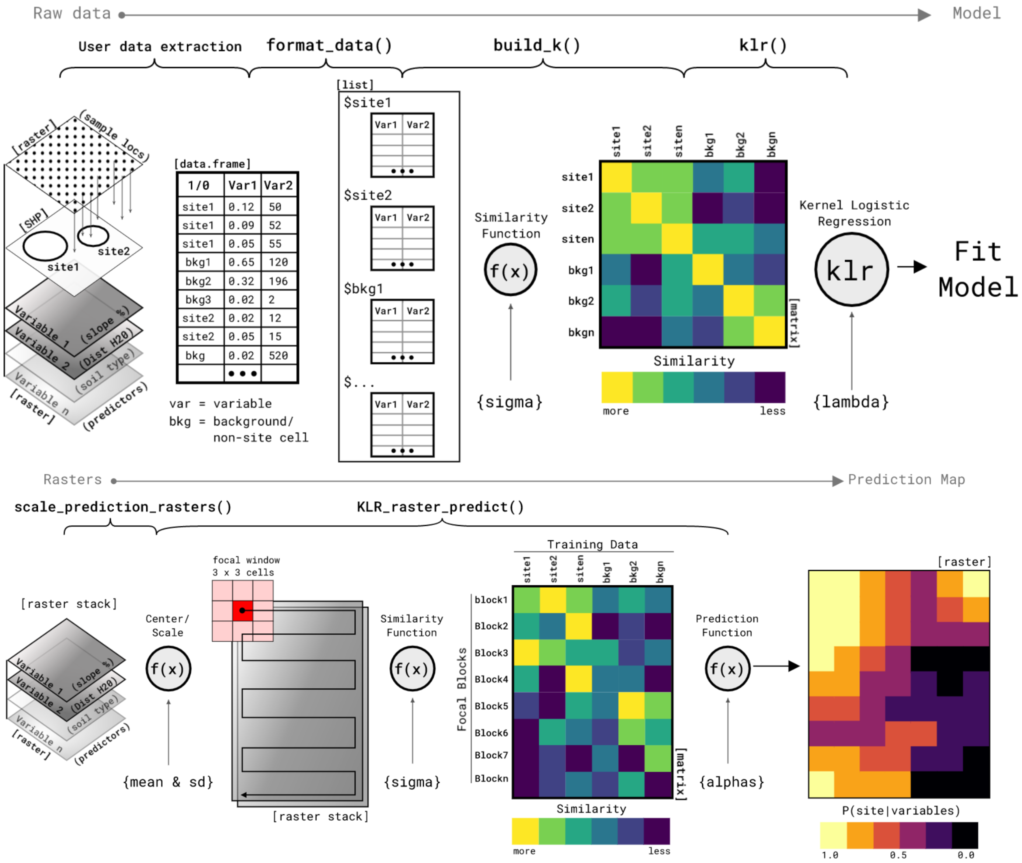
Summary

KLRfome is a package for predicting a geographic area's realtive sensitivity for the presence of archaeological sites. This is achieved by fitting, predicting, and visualizing a Kernel Logistic Regression model on mean feature embeddings. This regression algorithm and package are created to improve upon the current methods in archaeological predictive modeling. These improvements include 1) modeling rich descriptions of archaeological landforms by mitigating undesiarable spatial corraltion between samples, 2) explicitly modeling the similarity between archaeological sites as characterized by rich features, 3) the ability to define research specific similairty measures, and 4) focal window prediction that can be modified based on theory or managment goals.

More specifically, this package seeks to solve the *Distribution Regression* problem for archaeological site location modeling; or any other data for that matter. The aim of Distribution Regression is to map a single scalar outcome (e.g. presence/absence; 0/1) to a distribution of features. This is opposed to typical regression where you have one observation mapping a single outcome to a single set of features/predictors. For example, an archaeological site is singularly defined as either present or absent, however the area within the sites boundary is not singularly defined by any one measurement. The area with an archaeology site is defined by an infinite distribution of measurements. Modeling this in traditional terms means either collapsing that distribution to a single measurement or pretending that a site is actually a series of adjacent, but independent measurements. The methods developed for this package take a different view instead by modeling the distribution of measurements from within a single site on a scale of similarity to the distribution of measurements on other sites and the environmental background in general. This method avoids collapsing measurements and promotes the assumption of independence from within a site to between sites. By doing so, this approach models a richer description of the landscape in a more intuitive sense of similarity.

To achieve this goal, the package fits a Kernel Logistic Regression (KLR) model onto a mean embedding similarity matrix and predicts as a roving focal function of varying window size. The name of the package is derived from this approach; **K**ernel **L**ogistic **R**egression on **FO**cal **M**ean **E**mbeddings (**klrfome**) pronounced *clear foam*. This work is based on Szabó's work on feature mean embeddings for distribution regression (Szabó et al. 2016, Szabó et al. (2015)) and the Kernel Logistic Regression algortihm in (Zhu and Hastie 2005). A poster introducing the KLRfome package was presented on April 12, 2018 at the Society for American Archaeology annual meeting in Washington, D.C.The KLRfome package is available at https://github.com/mrecos/klrfome and DOI https://doi.org/10.5281/zenodo.1218403

The R package to impliment this KLRfome model contains functions formatting tabular data `format_data()` into the a specified list format including training and testing data splits, building a Guassian (RBF) similarity kernel `build_k()`, fitting the klr model `klr()`, and predicting the fit model to a stack of raster inputs `KLR_raster_predict()`. The hyperparameters of this model are `sigma` to control the variance of the Guassian kernel, `lambda` to control the regularization penalty of the regression model, and `ngb` to control the neighborhood dimensions for the focal prediction function. The also package contains functions to simulate data `get_sim_data()` and correlated trend rasters `sim_trend` (using the NLMR package (Sciaini, Fritsch, and Simpkins 2017)) that proxy archaeological site locations and environments so that the package can be tested on smaller and less sensitive data sets. Finally, with an appropriate parallel processing backend, such as `doParallel`, the `KLR_raster_predict()` function can be parameterized to split the prediction area into blocks (`split`, `ppside`, and `parallel` arguments) and process each block on a seperate core in parrallel. Each block will be collared by the focal window `ngb` size so that when merged back together there is no edge effect from predicting as blocks.



1/0	Var1	Var2
site1	0.12	50
site1	0.09	52
site1	0.05	55
bkg1	0.65	120
bkg2	0.32	196
bkg3	0.02	2
site2	0.02	12
site2	0.05	15
bkg	0.02	520
	•••	

var = variable
bkg = background/
      non-site cell

References

Sciaini, Marco, Matthias Fritsch, and Craig E. Simpkins. 2017. *NLMR: Simulating Neutral Landscape Models* (version 0.2.0). https://CRAN.R-project.org/package=NLMR.

Szabó, Zoltán, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. 2015. "Two-Stage Sampled Learning Theory on Distributions." In *International Conference on Artificial Intelligence and Statistics (Aistats)*, 948–57.

Szabó, Zoltán, Bharath Sriperumbudur, Barnabás Póczos, and Arthur Gretton. 2016. "Learning Theory for Distribution Regression." *Journal of Machine Learning Research* 17:1–40.

Zhu, Ji, and Trevor Hastie. 2005. "Kernel Logistic Regression and the Import-Vector Machine." *Journal of Computational and Graphical Statistics* 14 (1):185–205.