The Journal of Open Source Software

Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data

William D McGinnis[1, 2], **Chapman Siu**[3], **Andre S**[4], **and Hanyu Huang**[5]

1 Predikto, Inc. **2** Helton Tech, LLC **3** Suncorp Group Ltd. **4** Jungle AI **5** Tencent, Inc.

DOI: 10.21105/joss.00501

Software

- Review
- Repository
- Archive

Submitted: 05 December 2017
Published: 22 January 2018

Summary

Category_encoders is a scikit-learn-contrib module of transformers for encoding categorical data. As a scikit-learn-contrib module, category_encoders is fully compatible with the scikit-learn API (Buitinck et al. 2013). It also uses heavily the tools provided by scikit-learn (Pedregosa et al. 2011) itself, scipy(Jones et al. 2001–2001--), pandas(McKinney 2010), and statsmodels(Seabold and Perktold 2010).

Categorical variables (wiki) are those that represent a fixed number of possible values, rather than a continuous number. Each value assigns the measurement to one of those finite groups, or categories. They differ from ordinal variables in that the distance from one category to another ought to be equal regardless of the number of categories, as opposed to ordinal variables which have some intrinsic ordering. As an example:

Ordinal: low, medium, high Categorical: Georgia, Alabama, South Carolina, … , New York

The machine learning algorithms we will later use tend to want numbers, and not strings, as their inputs so we need some method of coding to convert them.

Category_encoders includes a number of pre-existing encoders that are commonly used, notably Ordinal, Hashing and OneHot encoders ("R Library Contrast Coding Systems for Categorical Variables," n.d.)(Carey 2003)(Weinberger et al. 2009). There are also some less frequently used encoders including Backward Difference, Helmert, Polynomial and Sum encoding ("R Library Contrast Coding Systems for Categorical Variables," n.d.)(Carey 2003). Finally there are experimental encoders: LeaveOneOut, Binary and BaseN (Zhang, n.d.)(McGinnis 2016a)(McGinnis 2016b).

The goal of these sorts of transforms is to represent categorical data, which has no true order, as numeric values while balancing desires to keep the representation in as few dimensions as possible. Category_encoders seeks to provide access to the many methodologies for accomplishing such tasks in a simple to use, well tested, and production ready package.

References

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, et al. 2013. "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project." *arXiv Preprint arXiv:1309.0238*.

Carey, Gregory. 2003. "Coding Categorical Variables (Http://Psych.colorado.edu/ Carey/Courses/

Jones, Eric, Travis Oliphant, Pearu Peterson, and others. 2001–2001--. "SciPy: Open Source Scientific Tools for Python." http://www.scipy.org/.

McGinnis, William D. 2016a. "Beyond One-Hot: An Exploration of Categorical Variables." *Will's Noise*. http://www.willmcginnis.com/2015/11/29/beyond-one-hot-an-exploration-of-categor

———. 2016b. *Will's Noise*. http://www.willmcginnis.com/2016/12/18/basen-encoding-grid-searc encoders/.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 51–56.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *The Journal of Machine Learning Research* 12. JMLR. org:2825–30.

"R Library Contrast Coding Systems for Categorical Variables." n.d. *IDRE Stats*. https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *9th Python in Science Conference*.

Weinberger, Kilian, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. "Feature Hashing for Large Scale Multitask Learning." *Proceedings of the 26th Annual International Conference on Machine Learning - ICML 09*. https://doi.org/10.1145/1553374.1553516.

Zhang, Owen. n.d. "Strategies to Encode Categorical Variables with Many Categories." *Caterpillar Tube Pricing | Kaggle*. https://www.kaggle.com/c/caterpillar-tube-pricing/discussion/15748#143154.