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

William D McGinnis¹, ², Chapman Siu³, Andre S⁴, and Hanyu Huang⁵

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

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


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


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