More About GLRM

There are a few ways a dataset can be “messy”: some values may be missing, some features may be categorical, numerical features may vary in orders of magnitude, or there may be outliers. Even trickier, one or more of the features of the dataset may be a sequence or time series.

Regardless of how such data could be represented (say, a relational database or a nested dictionary), we urge the reader to consider such data as a table, where rows correspond to examples and columns correspond to features. We describe each feature as having a “type”, such as a real-valued number, a category, or a rating from 1-7. Or even a time-series, such as a credit history. Each example is described by its features, which may be a mix of data types.

Our goal is to cluster examples from this table when some features are times series and the dataset is messy. Our approach is inspired by Generalized Low Rank Models (Udell ’16), and indeed can be viewed as an extension that accommodates the new “sequential” data type.

At Retina AI we use this approach to predict Customer Lifetime Value through segmentation of pyscho-graphic behavior, where features may be time-series (browsing pattern), a pmf (device propensity), categorical (acquisition channel), counts (visits) or numerical (total spend). We can tune our clustering models to be representative, descriptive or predictive based on the business use case. We’re currently developing a Python package for our work.

BACK

What's different about clustering this time

- Use Generalized Low Rank Model instead of traditional K-Means approach to tackle issues around outliers, missing data, categorical data and custom optimization functions
- Built at Stanford & Cornell University by Retina Team members
- Impute missing data vs. ignore the whole data point
- Simple REST API to score new customers
- Makes Segmentation Actionable in less than a day

NEXT