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
- Imputes missing data vs. ignoring the whole data point
- Current used by Madison Reed and Chegg to build omni-channel strategy
- Simple REST API to score new customers
- Makes Segmentation Actionable in less than a day

NEXT

User Stories from the field

Madison Reed Hair Color

- Using Character Design Tool to segment not only 2M customers but also over 10M prospects.
- Running Prospecting and Retargeting campaigns based on these personas to increase conversion rates by 45%
- Using the Early CLV value to understand behavior of the customer at time of conversion.
- Used by customer service team to real-time ping the Retina API to get (1) Persona (2) Residual LTV (3) Risk Scoring.

Chegg®

Chegg Study

- Building personas of students along with their propensity to truly churn.
- Using these clusters to run deeper surveys and focus groups to get understanding of who the customer is.
- Marketing team using personas to build personalized content for each persona.
- Using risk scoring to run retention campaigns reducing customer churn by 11%

NEXT

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