Categorical Embeddings: New Ways to Simplify Complex Data

rstudio::global(2020)

Alan Feder

Principal Data Scientist
Invesco

Two Types of Data

- **Continuous**: Numbers
  - e.g. 3, 6.5, -8.39, π
- **Categorical**: Categories
  - *Ordinal*: e.g. Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
  - *Nominal*: e.g. color, zip code

Modeling with tabular data

- e.g. from `readr::read_csv()` or `readxl::read_excel()`

Most statistical or supervised learning models implicitly assume that all your data is numeric

- e.g. Inverting matrices for linear regression or gradient descent for neural networks

Categorical data is common

- Election data may include state or zip code
- Marketing data may include profession or education level
- Financial data may include country or industry

Common ways of dealing with categories

- Ignore
    - But the data may be important!
- Label encode
    - Implies ordering, which might not be correct (e.g. is Red < Yellow < Blue???)
- Create dummy variables (i.e. one-hot encode)
    - Do you really want one column for each of 42,000 zip codes?

Categorical Embeddings

- Conceptually similar to word embeddings (e.g. Word2Vec or GloVe)
- Represented as a vector of numbers (e.g. length 2 or length 300)
- Each element of that vector represents something about the category itself

Benefits

- Represent all categories as a few numeric variables
- Learn more about your data with feature reduction (e.g. PCA or UMAP)
- Use them in your favorite models - even non-neural networks

Creating Embeddings

- Fit with neural networks
  - Can do it with `tensorflow`/`keras` or `torch`
- Now -- there is a tidymodels extension package: `embed`!

Conclusions

- Categorical embeddings are a useful way to use categorical data without creating dummy variables.
- The `embed::step_embed()` function is a great way of creating these embeddings in a tidy framework.
- You can then use these embeddings as inputs to a different model or by analyzing them directly (or both!)

Future areas of study

- How to decide on embedding length? (`num_terms =`)
- Hidden layers? (`hidden_units =`)
- Fit alongside other continuous predctors (`predictors =`)

Acknowledgements

- I first learned about categorical embeddings from fastai: Practical Deep Learning for Coders
- Data and analysis steps taken from TensorFlow training at RStudio::conf(2019) by Sigrid Keydana, Kevin Kuo, Rick Scavetta
- Slide format created with xaringan by Yihui Xie

Thank you

slides & code: https://github.com/AlanFeder/rstudio_2020_embed

contact: AlanFeder@gmail.com

Twitter: @AlanFeder