A Proposal to Research General Latent Feature Modeling

Anfeng Yu

September 1, 2018

1 Introduction

General Latent Feature Modeling (GLFM) [1] is a general unsupervised learning method, which can be applied to heterogeneous datasets (different variables can be either discrete or continuous).

This GLFM is a general method to get latent features (the hidden data structure inferred from observed features). And let me first explain what original latent feature models can do. A simple example is from Figure 1 to Figure 2. This is a simply generated database. The original features have been randomly combined and added noise, then generated a huge and redundant database with 36 new observable features. But with GLFM or other latent feature model, these four latent features still can be successfully recovered.

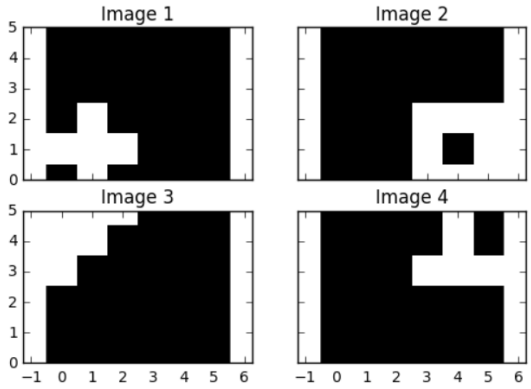

Figure 1: Original data's hidden features (prior distributions), from GLFM [1] github code

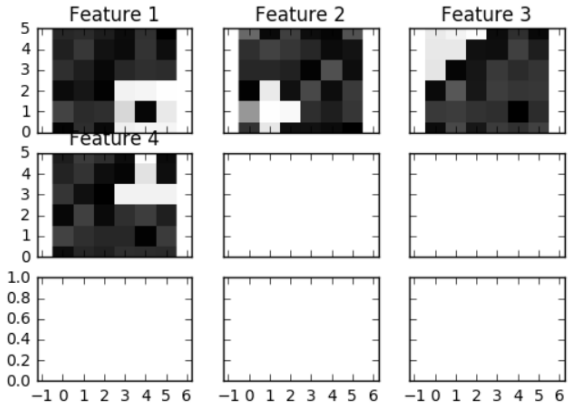

Figure 2: Inferred latent features by GLFM, from GLFM [1] github code.

2 GLFM's theory

GLFM is based on Indian Buffet Process model (IBP), which is a method to generate a prior latent feature matrix, as Figure 3. With this matrix, we can apply original data to a linear-Gaussian

latent feature model (or other Bayesian models). By optimization, finally we can change this generated prior latent feature matrix to true latent feature model.

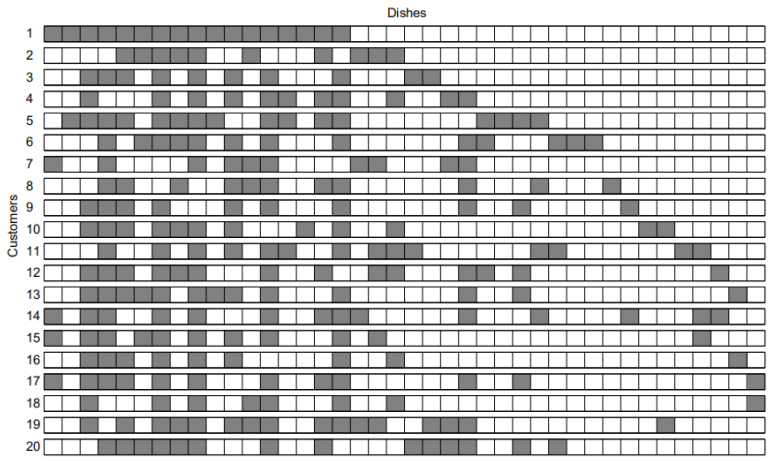

Figure 3: The way IBP generate latent feature matrices, from IBP [2]

Based on IBP, GLFM has two key improvements. First, GLFM introduced a pseudo-observation variable. With this pseudo-observation variable, GLFM can behave as standard linear-Gaussian IBP. Second, GLFM found a function that transforms the pseudo-observation into the actual observation. With these two key improvement, GLFM can be applied on heterogeneous datasets and work as a standard linear-Gaussian IBP.

3 Plan of Action

1.Using the GLFM and its given github code for Prostate Cancer dataset and Social Background on Mental Disorders dataset, then get the latent feature model and data analysis results (for example, the main reason for mental disorder) for each datasets.

2.Using other based on IBP latent feature models, like linear-Gaussian latent feature model with binary features, on same datasets to get their data analysis result.

3.Comparing two methods' data analysis results, to see if GLFM has a general and applicable interpreting ability.

4.For further research, applying GLFM to some heterogeneous datasets which traditional algorithms are hard to deal with (like IMDB movie data, which is heterogeneous and complicated).

4 Datasets For Research

1.(Used in the paper) Prostate Cancer dataset
http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

2.(Used in the paper) Social Background on Mental Disorders
https://aspe.hhs.gov/report/data-health-and-well-being-american-indians-alaska-natives-and-other-native-americans [3]

3.IMDB data
https://www.kaggle.com/PromptCloudHQ/imdb-data
https://www.imdb.com/interfaces/

4.Rotten Tomatoes data
https://developer.fandango.com/rotten_tomatoes

References

[1] I. Valera, M. F. Pradier, and Z. Ghahramani, "General Latent Feature Modeling for Data Exploration Tasks," *ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, 2017.

[2] T. L. Griffiths and Z. Ghahramani, "The Indian buffet process:an introduction and review," *Journal of Machine Learning Research*, 2011.

[3] F. J. R. Ruiz, I. Valera, C. Blanco, and F. Perez-Cruz, "Bayesian Nonparametric Comorbidity Analysis of Psychiatric Disorders," *Journal of Machine Learning Research.*, 2013.