Bayesian Nonparametrics and Application on Text Style Analysis

Jingze Zhang, Gordon Erlebacher
Department of Scientific Computing, Florida State University

Abstract

Over the last few years, NLP researchers have proposed various machine learning models to analyze, understand, and generate human language. While most existing methods often focus on semantic or word analysis, language style has been far less investigated. Moreover, extant research makes simplifying assumptions that classify styles coarsely using a few categories such as positive and negative sentiment. This study aims to develop an unsupervised nonparametric Bayesian method to identify different styles without a pre-specification of style categories. Nonparametric Bayesian methods can adapt their parameter dimensions depending on the data complexity to effectively and efficiently model the observed data. Its adaptability fits our style identification problem well where we do not know a priori the number of language styles in use, even by a single person. For example, we hypothesize that a person expressing deception will do so in a style different from when the writer is honest. In this work, we present a latent feature model based on the Indian Buffet Process (IBP). The nonparametric Bayesian method on style analysis is still under development.

Backgrounds

Natural language processing (NLP) is a field in which computers analyze, understand, and generate human language. Many NLP tasks focusing on semantics have been closely studied, such as entity recognition and text generation. As opposed to semantics, the area of language style is relatively uncultivated. Language style is quite rich since even a single person may have several ways to convey the same meaning. Common studies transfer text from one style to another with a simplifying assumption that classifies styles coarsely using a few categories such as positive and negative sentiment. This study aims to develop an unsupervised method to identify different styles without specifying the style categories.

Methods

▪ Bayesian Nonparametrics

Bayesian Nonparametrics is a branch of Bayesian analysis. In contrast to a parametric model in which we assume a fixed number of parameters, a nonparametric model has an unbounded number of parameters. It can adapt its potentially infinite degrees of freedom as more data accrue. Its adaptability lets it effectively and efficiently model data with arbitrary complexity and empowers continuous learning when new data emerge. To have infinite dimensions, Bayesian Nonparametric methods require stochastic processes as prior distributions.

▪ Indian Buffet Process

Indian Buffet Process (IBP) is one of the broadly-used stochastic processes in Bayesian Nonparametrics. It defines a probability distribution of binary sparse matrices with a finite number of rows and an infinite number of columns. The IBP is analogously defined as follows. Consider N customers sequentially select select dishes from an Indian buffet with a countably infinite number of dishes. Denote the customers' choices by a binary matrix Z, in which the rows are customers and columns are dishes. $Z_{ik} = 1$ if customer i takes dish k. The first customer starts from the left and continuously takes a $Poisson(\alpha)$ number of dishes. The i th customer chooses each previously-taken dish with probability $\sum_{j=1}^{i-1} Z_{jk} / i$, then tries a $Poisson(\alpha / i)$ number of new continuous dishes. A sample from IBP is shown in Figure 1. In reality, such a feature matrix is unknown and is inferred from data X using sampling methods or neural networks.

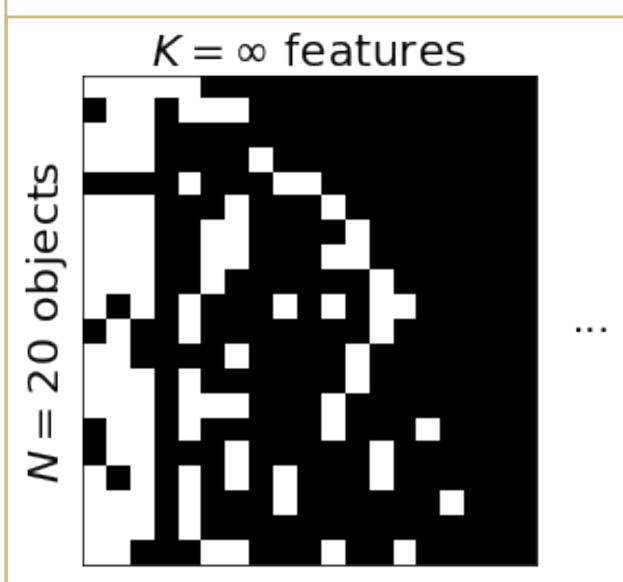

Fig 1. *A sample from IBP with 20 objects/people and the concentration parameter α = 5. The pixel is white if the object has the feature.*

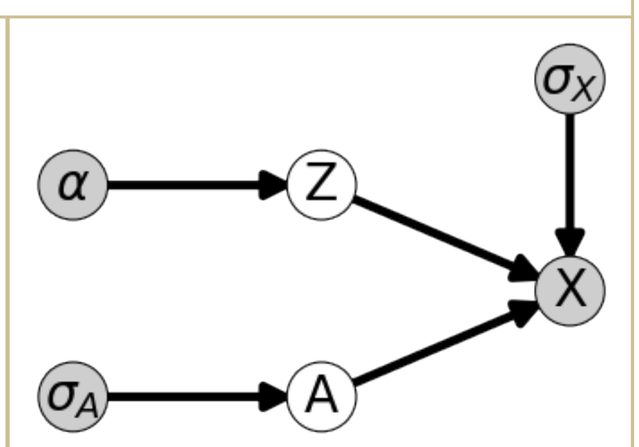

Fig 2. Graphical model of the latent feature model, where A is the weight matrix following a normal distribution and Z is the binary feature matrix from the Indian Buffet Process. X is the observed data

Experiments

▪ Data

We build a word collection by manually choosing five categories and ten words from each category. For each word, we look up its word embeddings from the Glove 50-dimension embedding dictionary and concatenate them together as the data X. The goal is to find the five categories. We also propose a semi-supervised method (semi-IBP) that can take known labels into account and infer the rest.

▪ A latent feature model

We assume the data X is generated following a linear gaussian latent feature model. Specifically,

$$X \sim N(Z \cdot A, \sigma_X)$$
$$Z \sim IBP(\alpha)$$
$$A \sim N(0, \sigma_A)$$

where A is the weight matrix that defines the features, and Z is the binary feature matrix that defines the features of each object. α is the prior concentration parameter. σ_X and σ_A set the prior diffuseness of X and A, respectively. A graphical model is shown in Figure 2.

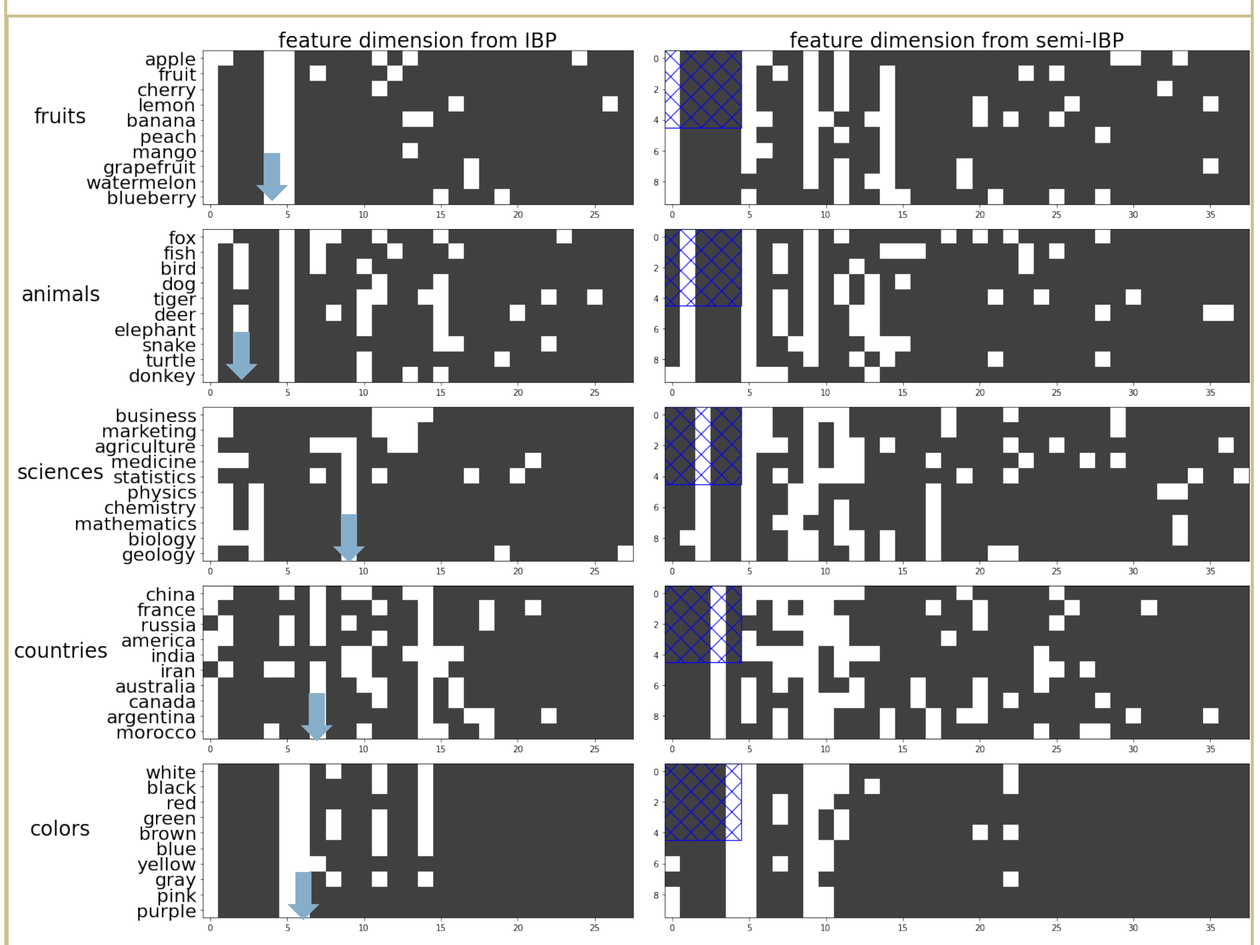

Fig 3. Left: the feature matrix learned from IBP and reordered by category. Category-corresponding column is labeled with a blue arrow. Right: the feature matrix learned from semi-IBP and reordered by category. The blue shadow indicates the frozen labels.

Figure 3 shows the feature matrices generated by IBP and semi-IBP. Rows are reordered by category for easier comparison. The blue shadow area in the right feature matrix represents the given labels to start semi-IBP. By providing labels for the first five words in each category, we also determine the first five columns representing fruits, animals, sciences, countries, and colors. As it was shown, semi-IBP can identify the remaining words. However, it takes more effort to judge whether standard IBP identifies the correct categories because of the exchangeable columns. The correct column for each category is labeled with a blue arrow. As shown in the table below, semi-IBP can identify desired features more accurately.

accuracy	fruit	animals	sciences	countries	colors	avg
IBP	96%	90%	88%	84%	100%	91.6%
Semi-IBP	84%	96%	100%	100%	100%	96%

Future works

Word embedding is considered carrying both semantic and style information. How to disentangle style information from semantic is an important and difficult problem towards style analysis.

Reference

[1] Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2019). Neural style transfer: A review. IEEE transactions on visualization and computer graphics, 26(11), 3365-3385.

[2] Griffiths, T. L., & Ghahramani, Z. (2011). The Indian Buffet Process: An Introduction and Review. Journal of Machine Learning Research, 12(4).

[3] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical