Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Bayesian Non-Parametrics

Advanced Machine Learning for NLP

Jordan Boyd-Graber

SLIDES ADAPTED FROM ELI BINGHAM AND MATT DICKENSON

Outline

1. Latent feature models
2. Finite latent **feature** (i.e. binary) models
3. (Very fast introduction)
4. Application: Topic Models

Latent feature models

- Feature model: N items described by K features
- Dense feature model: every feature is present in every item, e.g. PCA
- Sparse feature model: only some features present in each item, and we can assume feature values and presence are independent:

$$
\begin{aligned}
\mathbf{X} &= \mathbf{A} \otimes \mathbf{Z} \\
P(\mathbf{X}) &= P(\mathbf{A}) P(\mathbf{Z})
\end{aligned}
$$

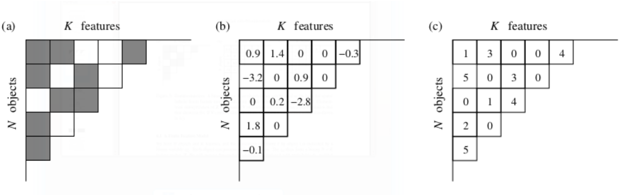


Figure: Griffiths and Ghahramani (2011) Figure 3

Example

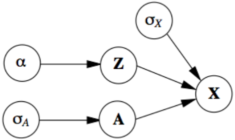
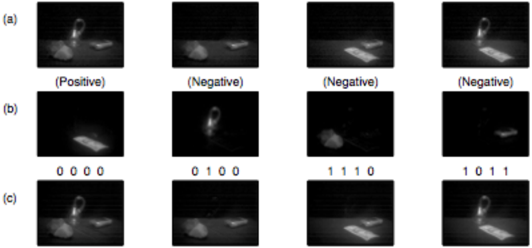


Figure: Griffiths and Ghahramani (2011) Figure 7



Motivation

- Problem with finite latent feature model: K is fixed
- Goal: construct nonparametric prior on $\mathbf{Z}$ so that K grows with the complexity of the dataset
- As with DPMMs, we can try to build one by taking $K \to \infty$ in a finite feature model

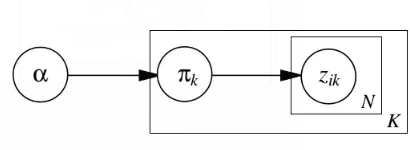


Figure: Griffiths and Ghahramani (2011) Figure 4

Finite latent feature models

The basic finite distribution on $z_{i,k}$s:

$$\begin{aligned} \pi_k|\alpha &\sim \text{Beta}\left(\frac{\alpha}{K}, 1\right) \\ z_{i,k}|\pi_k &\sim \text{Bernoulli}(\pi_k) \end{aligned}$$

As with DPMMs, we can marginalize out latent feature presence probabilities π_k to obtain a distribution on matrices $\mathbf{Z} \in \{0,1\}^{N \times K}$:

$$\begin{aligned} P(\mathbf{Z}) &= \prod_{k=1}^{K} \int \left(\prod_{i=1}^{N} P(z_{ik}|\pi_k) \right) P(\pi_k) d\pi_k \\ &= \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})} \end{aligned}$$

The $K \to \infty$ limit

- lof($\mathbf{Z}$) is the matrix obtained by ordering the columns of $\mathbf{Z}$ as N-digit binary numbers
- To define a probability over infinitely wide binary matrices using de Finetti's Theorem, we need exchangeable symmetry, so we define *lof* equivalence classes by modding out column order:

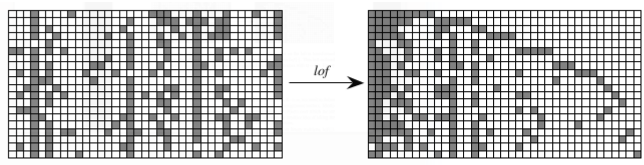


Figure: Griffiths and Ghahramani (2011) Figure 5

Indian Buffet Process:

1. N customers enter (in sequence) a buffet restaurant with an infinite number of dishes
2. First customer fills her plate with Poisson(α) number of dishes
3. i^{th} customer samples dishes in proportion to their popularity, with probability $\frac{m_k}{i}$, where m_k is the number of previous customers who sampled dish k
4. i^{th} customer then samples $K_1^{(i)} \sim$ Poisson($\frac{\alpha}{i}$) number of new dishes

Resulting probability distribution on matrices:

$$P(\mathbf{Z}) = \frac{\alpha^{K_+}}{\prod_{i=1}^{N} K_1^{(i)}!} \exp(\alpha H_N) \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

Alternative derivation: Stick-Breaking

1. Recursively break (an initially unit-length) stick, breaking off a Beta$(\alpha, 1)$ portion at each step
2. Let each portion of the "stick", π_k represent the probability of each feature (sorted from largest to smallest)

This helps to show the relation between the Dirichlet process and the IBP. The stick-breaking construction is also useful for defining inference algorithms.

Application: Topic Modeling

"The IBP Compound Dirichlet Process
and its Application to Focused Topic Modeling"
Williamson, Wang, Heller, and Blei (2010)

Stick-breaking construction:

$$\mu_k \sim \text{Beta}(\alpha, 1)$$

$$\pi_k = \prod_{j=1}^{k} \mu_j$$

$$b_{m,k} \sim \text{Bernoulli}(\pi_k)$$

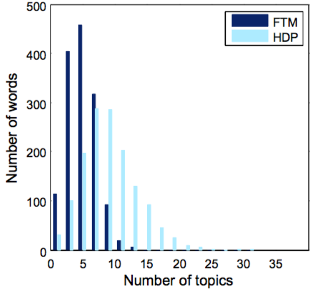
Application: Topic Modeling

Focused topic model:

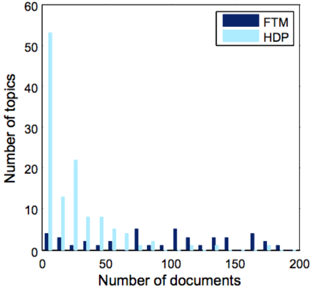
1. for $k = 1, 2, \ldots$
   - Sample stick length π_k
   - Sample relative mass $\phi_k \sim \text{Gamma}(\gamma, 1)$
   - Draw topic distribution over words: $\beta_k \sim \text{Dirichlet}(\eta)$
2. for $m = 1, \ldots, M$
   - Sample binary vector b_m
   - Draw total number of words $n^{(m)} \sim NB(\sum_k b_{m,k}\phi_k, 1/2)$
   - Sample distribution over topics $\theta_m \sim \text{Dirichlet}(b_m \cdot \phi)$
   - For each word $w_{m,i}, i = 1, \ldots, n^{(m)}$
     1. Draw topic index $z_{m,i} \sim \text{Discrete}(\theta_m)$
     2. Draw word $w_{m,i} \sim \text{Discrete}(\beta_{z_{m_i}})$

Application: Topic Modeling

Number of Topics a Word Appears in



Number of Documents a Topic Appears in



- Separates global topic proportions from per-document distribution
- Rare topics can dominate documents
- Frequent topics can't appear in as many documents

Discussion

Limitations of IBP:

1. Coupling of average number of features α and total number of features $N\alpha$ (can be overcome with a two-parameter generalization)
2. Computationally complex, can be time-consuming

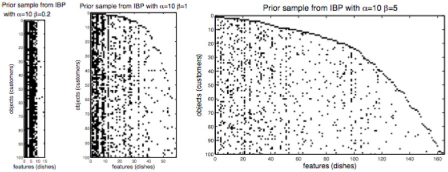


Figure: Griffiths and Ghahramani (2011) Figure 10

Connection to Rest of Course

+ Bayesian Nonparameterics discovers dimension
+ Strong probabilistic foundations
+ Gives meaning to representation
- Hard to implement
- Slow
- Not as effective