UNIVERSITY OF CAMBRIDGE

Correlated Non-Parametric Latent Feature Models

Finale Doshi-Velez and Zoubin Ghahramani, University of Cambridge

Abstract

We are often interested in explaining data through a set of hidden factors or features. When the number of hidden features is unknown, the Indian Buffet Process (IBP) is a nonparametric latent feature model that does not bound the number of active features in dataset. However, the IBP assumes that all latent features are uncorrelated, making it inadequate for many realworld problems. We introduce a framework for correlated nonparametric feature models, generalising the IBP. We use this framework to generate several specific models and demonstrate applications on realworld datasets.

Motivation

Observations can often be explained by a set of features, but we don't know a priori how many features there are.

computer
keyboard
mouse

In many cases, we expect these features to be correlated:

computers
keyboards
mice

knives
spoons
forks

We propose a general framework for creating a nonparametric correlated feature model.

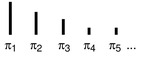
The Indian Buffet Process

The Indian Buffet Process (IBP) is a non-parametric prior on binary matrices—generally useful for latent feature models. The generative process proceeds as follows:

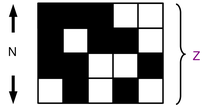
(1) We sample an infinite number of Beta(α,1) variables $v_1, v_2, v_3 \ldots$

(2) The probability of a feature k π_k is given by $\pi_k = \prod_{i=1}^{k} v_i$

Each feature has some observation parameters a_k associated with it.



(3) Each observation samples features based on their probabilities



Properties

- Observations are exchangeable.
- Features are independent.
- Finite number of features in a finite dataset.

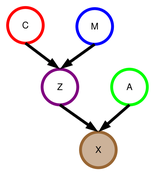
Aside: The Dirichlet Process (DP) is a distribution on distributions, where the probability of a category a_k is given by $\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$

General Framework

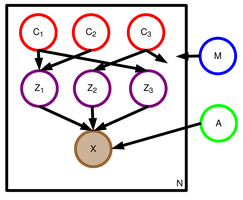
Desiderata for a correlated non-parametric feature model:

- A finite dataset should contain a finite number of latent features with probability one.
- Features and data should remain exchangeable.
- Correlations should capture motifs, or commonly occurring sets of features.

Our approach: introduce correlations through a hierarchical structure.

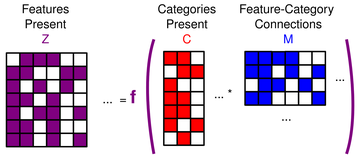


High level graphical model: correlations are introduced in the feature-assignment matrix Z by adding a layer of hierarchy: Z now depends on two other variables C and M. As in the IBP model, the features Z and some parameters A define the distribution on the data X.



We can think of C as a set of category -assignments that determines what higher-level categories observation n contains. The connection-assignment variable M defines how categories relate the the hidden features Z.

For C and M, we consider binary matrices C and M which relate to Z through matrix product:



Sufficient conditions on f, C, and M to satisfy the desiderata:

- **C** is generated by some nonparametric process NP1 that associates each observation with a finite number of categories with probability one.
- **M** is generated by some nonparametric process NP2 that associates each category with a finite number of features with probability one.
- **f** is a (potentially stochastic) link function such that
  - $z_{nk} = f(c_n^T m_k)$
  - if $c_n^T m_k = 0$ then $z_{nk} = 0$ with probability one.

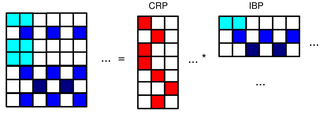
Specific Variants

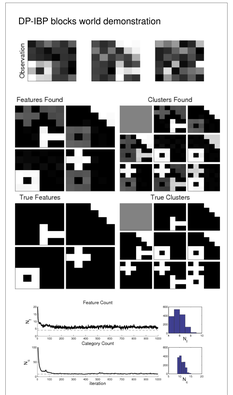
The **DP-IBP** is a structured clustering model in which each observation belongs to a single cluster, but features can be shared between clusters:

- C ~ CRP
- M ~ IBP
- $z_{nk} = c_n^T m_k$

The expected number of features grows as O(log(log(N)) as the number of observations grows.

Demonstration: we show a toy blocks world example that contains four features (see true features) that come in particular groupings (see true clusters). The inference recovers the clusters and features.





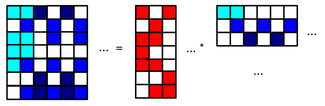
The **IBP-IBP** allows each observation to belong to multiple categories. Thus, observations consist of sets of sets:

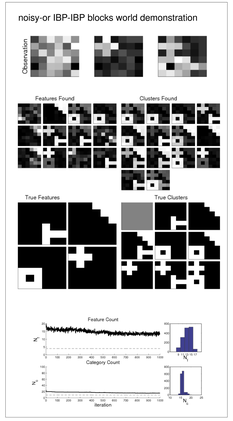
- C ~ IBP
- M ~ IBP
- $z_{nk} = (c_n^T m_k) > 0$

The **noisy-or IBP-IBP** extends the IBP-IBP by allowing features to be absent even if their parent categories are active:

- C ~ IBP
- M ~ IBP
- $z_{nk} \sim \text{Bernoulli}\,(1 - q^{c_n^T m_k})$

Demonstration: we show a similar blocks world example, except now the data was generated using an underlying noisy-or IBP-IBP model. Inference is more difficult and the data is more complex in this model, but the features and clusters are still largely recovered.





Experiments on Realworld Data

We compared the 3 structured models (DP-IBP, IBP-IBP, noisy-or IBP-IBP) to the DP and the IBP on five realworld datasets:

- UN: 15 development statistics from 155 countries
- India: 14 socioeconomic statistics from 398 households
- Joke: 500 user ratings of 30 jokes
- Gene: 226 gene expression levels from 251 subjects
- Robot: 23 tags for 750 images taken from a robot-mounted camera

The UN, India, Joke, and Gene datasets contained non-negative real-valued values. We used an exponential prior on A and assumed that the data had some additional Gaussian noise ε:

$$X = ZA + \varepsilon$$
$$A \sim \text{Exponential}(\lambda)$$
$$\varepsilon \sim N(0, \sigma_n)$$

The robot data was binary-valued. We used the Bernoulli likelihood:

$$P(X_{nd} = 1 \mid Z_{nd} = 1) = 1 - m$$
$$P(X_{nd} = 1 \mid Z_{nd} = 0) = f$$

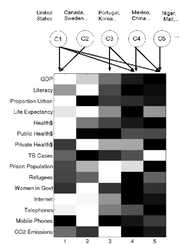
Quantitative results: overall, the simple DP mixture often had the best performance. The DP-IBP generally had the best performance among the features based models (including the IBP).

Qualitative results: the DP-IBP discovered interesting structures that the flat DP or IBP models could not.

Typical sample from robot data

- C1: hallway, door, trash can, chair, desk, office, computer, whiteboard
- C2: door, trash can, robot, bike, printer, couch
- C3: trash can, monitor, keyboard, book, robot, pen, plant
- C4: hallway, door, trash can

Typical sample from UN data



Conclusions

- We presented a framework for designing correlated nonparametric feature models using hierarchies.
- These models describe deep structures in the data, which may provide interesting explanatory power beyond what can be described by flat models like the DP.
- More powerful inference and tailored likelihood functions are needed to apply these models to realworld applications.