Variational Inference for Beta-Bernoulli Dirichlet Process Mixture Models

Mengrui Ni, Erik B. Sudderth (Advisor), Mike Hughes (Advisor)
Department of Computer Science, Brown University
mni@cs.brown.edu, sudderth@cs.brown.edu, mhughes@cs.brown.edu

Abstract

A commonly used paradigm in diverse application areas is to assume that an observed set of individual binary features is generated from a Bernoulli distribution with probabilities varying according to a Beta distribution. In this paper, we present our nonparametric variational inference algorithm for the Beta-Bernoulli observation model. Our primary focus is clustering discrete binary data using the Dirichlet process (DP) mixture model.

1 Introduction

In many biological and vision studies, the presences and absences of a certain set of engineered attributes create binary outcomes which can be used to describe and distinguish individual objects. Observations of such discrete binary feature data are usually assumed to be generated from mixtures of Bernoulli distributions with probability parameters $\boldsymbol{\mu}$ governed by Beta distributions. Such a model, also known as latent class analysis [6], is practically important in its own right. Our discussion will focus on clustering discrete binary data via the Dirichlet process (DP) mixture model.

2 Beta-Bernoulli Observation Model

The Beta-Bernoulli model is one of the simplest Bayesian models with Bernoulli likelihood $p(X \mid \mu)$ and its conjugate Beta prior $p(\mu)$.

2.1 Bernoulli distribution

Consider a binary dataset with N observations and D attributes:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix},$$

where $X_1, X_2, \ldots, X_N \sim$ i.i.d. Bernoulli$(X \mid \mu)$. Each X_i is a vector of 0's and 1's.

For a single Bernoulli distribution, the probability density for each observation has the following form:

$$p(X_i \mid \mu) = \prod_{d=1}^{D} \mu_d^{x_{id}} (1 - \mu_d)^{1 - x_{id}}. \tag{1}$$

Now, let's consider a mixture of K Bernoulli distributions. Each component has weight π_k, where $\sum_{k=1}^{K} \pi_k = 1$. Then, the density function for each observation can be expressed as follows:

$$p(X_i \mid \mu, \pi) = \sum_{k=1}^{K} \pi_k \, p(X_i \mid \mu_k), \tag{2}$$

where $\mu = [\mu_1, \ldots, \mu_K], \pi = [\pi_1, \ldots, \pi_K]$, and $p(X_i \mid \mu_k) = \prod_{d=1}^{D} \mu_{kd}^{x_{id}} (1 - \mu_{kd})^{1 - x_{id}}$.

2.2 Beta distribution

The conjugate prior to the Bernoulli likelihood is the Beta distribution. A Beta distribution is parameterized by two hyperparameters β_0 and β_1, which are positive real numbers. The distribution has support over the interval $[0, 1]$, and is defined as follows:

$$p(\mu \mid \beta_1, \beta_0) = \frac{1}{B(\beta_1, \beta_0)} \; \mu^{\beta_1 - 1} (1 - \mu)^{\beta_0 - 1}, \qquad \mu \in [0, 1]. \tag{3}$$

where $B(\beta_1, \beta_0) = \int_0^1 \mu^{\beta_1 - 1} (1 - \mu)^{\beta_0 - 1} \, d\mu$ is the beta function, which also has the form,

$$B(\beta_1, \beta_0) = \frac{\Gamma(\beta_1)\Gamma(\beta_0)}{\Gamma(\beta_1 + \beta_0)}, \quad \beta_0, \beta_1 > 0. \tag{4}$$

The hyperparameters β_0 and β_1 are "pseudocounts" which correspond to prior beliefs of the data before starting the experiment. Figure 1 shows Beta distributions for different values of β_0, β_1.

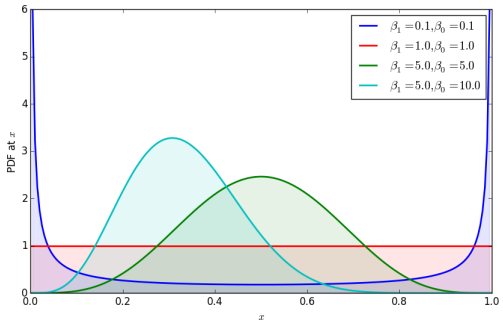

Figure 1: Examples of Beta distribution for different choices of hyperparameters

In general, we would like to set both β_0 and β_1 to be less than 1, in order to encourage sparse behavior within the sampled priors. In other words, by setting $\beta_0, \beta_1 < 1$, the sampled probability parameters μ should biased toward 0's and 1's, which produces very distinct priors that are useful for clustering. However, for some small datasets, Beta priors being too small could lead to bad local optima, which we will discuss in detail in the last section.

2.3 Bernoulli mixture model

We now return to our discussion of the Bernoulli mixture model, illustrated by the graphical model in Figure 2. For each observation X_i, we have a corresponding K-dimensional latent variable Z_i with a single component equal to 1 and the rest equal to 0. So, the conditional distribution of the latent variable Z given the weights is in the form:

$$p(Z \mid \pi) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}}. \tag{5}$$

Similarly, the Bernoulli likelihood of the observed data can be written as:

$$p(X \mid Z, \mu) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(X_i \mid \mu_k)^{z_{ik}}. \tag{6}$$

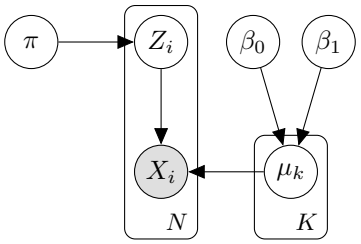

Figure 2: Graphical model representation of the Beta-Bernoulli mixture model. X_i's are observations which are governed by the Bernoulli distribution parameter μ. The latent variable Z_i is a K-dimensional vector with a single component equal to 1 and the rest equal to 0, indicating cluster assignment of X_i. β_0, β_1 are hyperparameters of the Beta distribution. π indicates the proportional of the components.

2.4 Dirichlet process (DP) mixture model

The Dirichlet process (DP) provides a nonparametric prior for partitioning exchangeable datasets into discrete clusters [3]. It is parameterized by a base distribution G_0 and a positive concentration parameter α. Then, $G \sim DP(\alpha, G_0)$ is a distribution over probability distributions, with stick-breaking weights $\{\pi_k\}_{k=1}^{\infty}$ sampled as follows:

$$G \triangleq \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \qquad v_k = \text{Beta}(1, \alpha), \qquad \pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell). \tag{7}$$

Each instance then draws an assignment Z_i from a categorical distribution of π, and an observation X_i from the Bernoulli distribution with parameter μ indicated by assignment Z_i. Here, we can also see the role of the concentration parameter α, namely higher values of α allow π_k to decay slower as $k \to \infty$ and thus encourage more clusters.

3 Variational Inference

Variational inference methods are well developed to provide solutions for mixtures of exponential families. Due to the intractability of working with the true posterior distribution, variational methods formulate the computation in terms of an optimization problem, which tries to approximate the posterior distribution of latent variables with a tractable family and then maximizes a low-bound of data likelihood with respect to variational parameters.

Specially, we define a variational distribution $q(Z, \mu, \pi)$ to approximate the true posterior distribution $p(Z, \mu, \pi|X)$. To do this, we minimize the Kullback-Leibler (KL) divergence with respect to the posterior distribution p:

$$q^*(Z, \mu, \pi) = \arg\min_q D(q||p) \tag{8}$$

The minimization in Equation (8) can be cast alternatively as the maximization of the evidence lower bound (ELBO) on the log marginal likelihood:

$$\log p(X \mid \beta_0, \beta_1, \alpha) \geq \mathbb{E}_q[\log p(Z, X, \mu, \pi \mid \beta_0, \beta_1, \alpha)] - \mathbb{E}_q[\log q(Z, \mu, \pi)] \triangleq \mathcal{L}(q) \tag{9}$$

Similar to the EM algorithm, the variational inference method alternates between local and global steps to update hidden variables until the evidence lower bound (ELBO) converges. The local variables Z_i only govern the distributions of their respective observation X_i. The global variables μ, π govern mixture components and mixture proportions. This algorithm is guaranteed to find a local optimum for ELBO.

Now, in order to avoid the algorithm converges to poor local optima, we also provide a memoized online variational algorithm that divides the data into predefined B batches $\{\mathcal{B}\}_{b=1}^{B}$. At each batch, we perform both the local and global updates. In addition, we develop two moves, "birth" and "merge". The "birth"

move can add useful new components to the model and escape poor local optima. The "merge" move, on the other hand, combines two components into a single merged one. In one pass of the data, the memoized algorithm performs a "birth" move for all batches, and several "merge" moves after the final batch.

4 Experiments

In this section, we evaluate the performance our model on a wide range of both synthetic and real-world binary datasets. We first generate synthetic data to explore the model's ability to infer the latent structures from observations. The dataset consists of 6 attributes and 5 true clusters with each cluster having 20 instances. Given the Bernoulli parameter below, we sample the $N = 100$ instances and run from 10 random initializations with $K = 5$:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} = \begin{bmatrix} .95 & .95 & .95 & .95 & .95 & .95 \\ .05 & .05 & .05 & .05 & .95 & .95 \\ .95 & .05 & .05 & .95 & .95 & .95 \\ .05 & .05 & .05 & .05 & .05 & .05 \\ .95 & .95 & .95 & .95 & .05 & .05 \end{bmatrix}.$$

Figure 3 shows that our variational algorithm successfully recovers the 5 true clusters. Each instance is also assigned to the correct corresponding cluster.

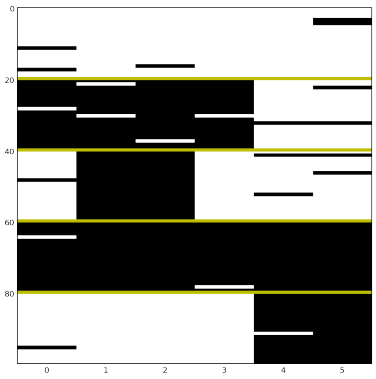

Figure 3: Synthetic binary data with 6 attributes and 5 true clusters with each cluster having 20 instances. White blocks have value "1" and black blocks have value "0". The yellow lines indicate separations of the clusters.

4.1 Animals with Attributes dataset

For their studies on attribute-based object similarity, Osherson and Wilkie collected judgements from human subjects on the "relative strength of association" between 85 attributes and 48 animal classes. Kemp et al. made use of the same data in a machine learning context and added 2 more animals classes. Figure 4 illustrates the resulting 50×85 class-attribute matrix.

We run our model with a Beta prior $([0.1, \ldots, 0.1], [0.1, \ldots, 0.1])$ that bias toward 0's and 1's and is uniform among features. Figure 4 shows one of the best runs from 20 k-means initializations with initial $K = 50$. We can see that the final number of components has merged down to 12. Importantly, the model also discovers relationships among animals where we have clusters of aquatic mammals, large tough-skin animals, and so on. Such partitions are almost indistinguishable from a human's attempt to cluster the dataset.

4.2 MNIST Handwritten Digits Database

We now test our model on the classical MNIST dataset with all $N = 60{,}000$ images of handwritten digits 0-9. Each image has $28 \times 28 = 784$ pixel features which are converted from real numbers to binary with threshold

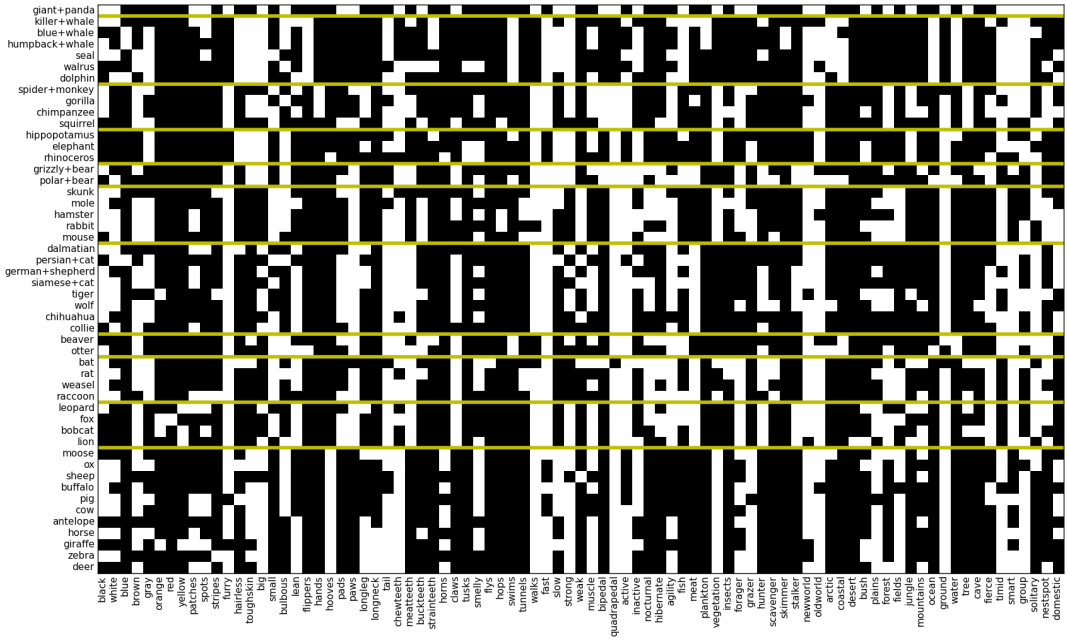

Figure 4: Animals with Attributes dataset with 50 animals and 85 binary attributes. Black blocks indicate 0 entries or absences of the attributes. White blocks indicate 1 entries or presences of the attributes. Yellow line indicates the separations of the animals. In this example, the animals are clustered into 12 groups.

on 150. Here we set the Beta prior $([0.1, \ldots, 0.1], [0.1, \ldots, 0.1])$ the same as before and runs from 10 k-means initializations with initial $K = 100$. One of the best runs eventually converges to 132 final components. Figure 5 shows the cluster means learned by the variational algorithm, which cover various styles of all the digits.

4.3 SUN Attributes Dataset

In our last experiment, we next learn the large-scale scene attribute database that build on top of the fine-grained SUN categorical database. The attribute database contains more than 14,000 images and 102 binary attributes related to materials, surface properties, lighting, functions and affordances, and spatial envelope properties. The dataset is again motivated by interests in "attribute-based" representation of visual objects as a complement to the "category-based" representations. Our goal here is to learn possible relationships between scene categories and attributes.

We again initialize our model with Beta prior $([0.1, \ldots, 0.1], [0.1, \ldots, 0.1])$ and concentration parameter $\alpha = 100$ to encourage more clusters. We run the model from 10 k-means initialization with initial $K = 100$. The final numbers of components merge to around 60. Figure 6 shows 10 image clusters with 10 randomly selected images from each cluster and a word cloud visualization showing the categories of the images. We can see that the attributes-based learning enable a higher understanding of the images which is difficult for simple scene recognition to achieve. For example, we observed images from sports events are clustered into the same group. Similarly, transportation terminals of flight and bus are in the same cluster. One can argue that the learned clusters are strongly affected by the particular set of attributes in the dataset. But, we can see that our model clearly learned the possible relationships between scenes in the images and cluster them in a logical way.

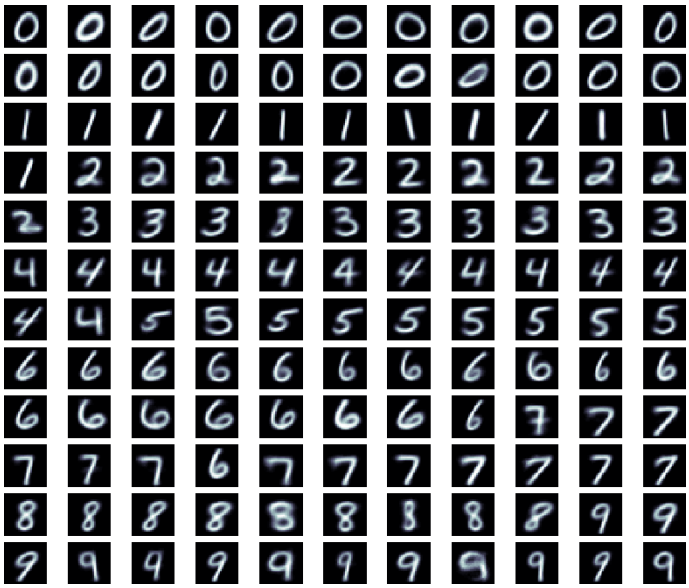

Figure 5: Visualization of the cluster means of the MNIST handwritten digits dataset. In this particular example, there are 132 components which represents different configurations of all the digits.

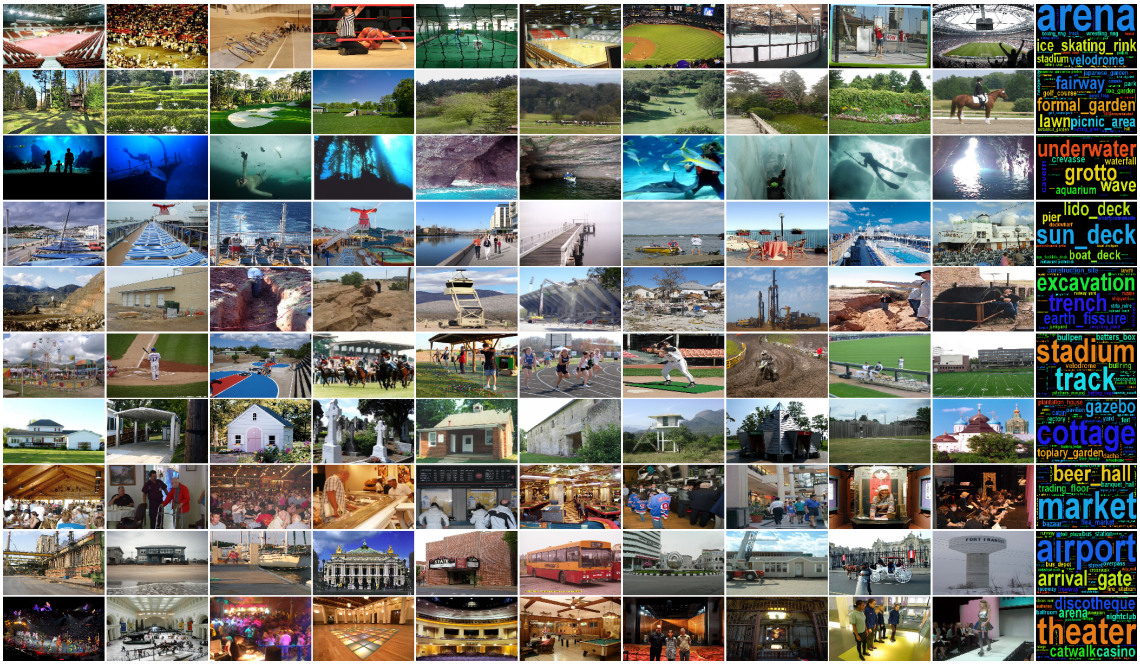

Figure 6: SUN Attributes dataset. Each row represents an image cluster. The columns are 10 randomly selected images from each corresponding cluster and a word cloud that visualize the scene categories of the images within the cluster. The font size of the words are proportional to the number of occurrences of images from the particular scene categories.

5 Discussions and Conclusions

We presented the Beta-Bernoulli observation model and a nonparametric framework for simultaneously clustering one or more sets of entities. The experiments above show our model successfully learned the latent structures of the binary mixture data.

It worths noting that choice of hyperparameters, in particular the Beta priors being $(0.1, 0.1)$, are not trivial. In our sensitivity analysis of the hyperparameters, we can see from Figure 7 that Beta priors equal to 0.1 gives the highest values of ELBO for both MNIST and SUN Attributes datasets. In general, β_1, β_0 being too big produces more uniform Beta priors among different clusters, which results in fewer clusters. For small values of λ, however, the result can be very poor if the size of the dataset is small. In particular, the Animals dataset with only 50 instances is extremely sensitive to the value of λ.

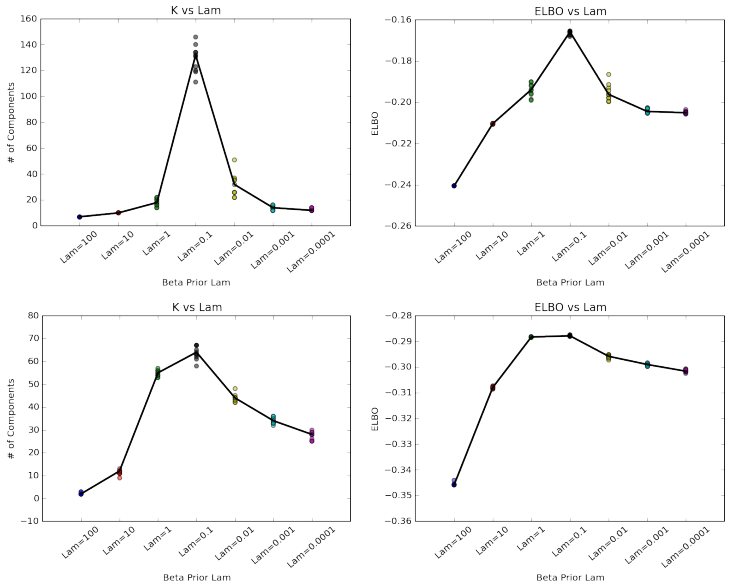

Figure 7: Comparison of the number of final components and ELBO versus λ. *Upper left:* Number of final components versus λ for the MNIST dataset. *Upper right:* Converged evidence lower bounds (ELBO) versus λ for the MNIST dataset. *Lower left:* Number of final components versus λ for the SUN Attributes dataset. *Lower right:* Converged evidence lower bounds (ELBO) versus λ for the SUN Attributes dataset.

References

[1] Blei, David M., and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." *Bayesian analysis* 1.1 (2006): 121-143.

[2] Genevieve Patterson, James Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. *Proceedings of CVPR* (2012).

[3] Hughes, Michael C., and Erik Sudderth. "Memoized online variational inference for Dirichlet process mixture models." *Advances in Neural Information Processing Systems*. (2013): (pp. 1133-1141).

[4] Yu, Shipeng, et al. "Variational bayesian dirichlet-multinomial allocation for exponential family mixtures." *Machine Learning: ECML 2006*. Springer Berlin Heidelberg, (2006). 841-848.

[5] Wang, Chong, John W. Paisley, and David M. Blei. "Online variational inference for the hierarchical Dirichlet process." *International conference on artificial intelligence and statistics*. (2011).

[6] McLachlan, G.J. and D. Peel. *Finite Mixture Models*. Wiley. (2000)

[7] Hughes, Michael C., Dae Il Kim, and Erik B. Sudderth. "Reliable and Scalable Variational Inference for the Hierarchical Dirichlet Process." *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. (2015).