Infinite Exponential Family Harmoniums

Ning Chen[†] **Jun Zhu**[‡] **Fuchun Sun**[†]
Department of CS & T, Tsinghua University, Beijing 100084, China
[†]{chenn07@mails, fcsun@mail}.tsinghua.edu.cn [‡]{jjzhunet9@hotmail.com}

Abstract

We present the *infinite exponential family Harmonium model* (iEFH), which is a bipartite latent feature model with an unbounded number of latent units. We use the Indian Buffet Process as a prior to enforce that the expected number of latent states to be finite with probability one. To the best of our knowledge, iEFH presents a first successful attempt towards utilizing the benefits of Bayesian nonparametrics to learn the structures of an undirected latent variable model.

1 Introduction

One challenging problem with learning both supervised and unsupervised latent feature models is to specify the number of latent units, which is a model selection problem. The development of Bayesian nonparametric methods that can automatically resolve the number of features from an infinite number of candidates a priori has the potential to bypass the model selection problem [2]. However, very few attempts have been made towards learning nonparametric undirected Markov networks such as harmoniums that have nice properties of fast inference due to the weak assumption that observations and response variables are conditionally independent given a set of latent variables. In this paper, we propose the *infinite exponential family Harmonium* (iEFH), which is to our knowledge the first attempt to learn infinite undirected latent variable models. We apply the nonparametric Indian Buffet Process (IBP) prior on an unbounded number of latent units but expect to select a finite subset when trained with a finite number of examples. We also extend iEFH to multi-view iEFH for incorporating heterogeneous data and to infinite max-margin Harmonium (iMMH) for achieving more accurate prediction results (e.g., classification accuracy). Finally, we extensively evaluate the proposed model on two real datasets and compare with several baseline methods.

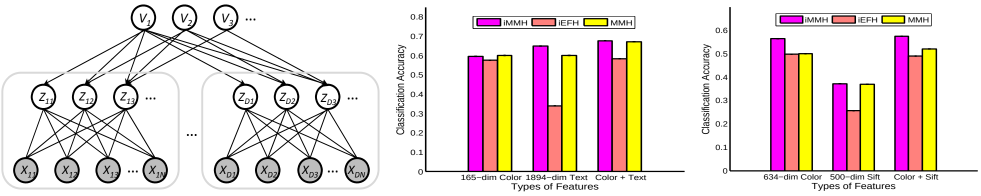
2 Infinite Exponential Family Harmoniums

The Model: Figure 1 (a) shows the unsupervised iEFH. The joint distribution is defined as: $p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\}) = p(\mathbf{v}) \prod_d p(\mathbf{z}_d|\mathbf{v}) p(\mathbf{x}_d|\mathbf{z}_d) \propto p(\mathbf{v}) \prod_d p(\mathbf{z}_d|\mathbf{v}) \exp\{\boldsymbol{\alpha}^\top \mathbf{x}_d - \frac{1}{2}\mathbf{x}_d^\top \mathbf{x}_d + \boldsymbol{\beta}^\top \mathbf{z}_d + \mathbf{x}_d^\top \mathbf{W} \mathbf{z}_d\}$. Let Z denote the random $D \times K$ binary matrix, where D denotes the number of observations and $K (\to \infty)$ denotes the number of latent features. Please refer to [3] for real-valued latent feature case. Each binary element $z_{dk} = 1$ if feature k is possessed by observation d, and 0 otherwise. Let $\pi_k \in (0, 1)$ be a parameter associated with each column of the binary matrix Z. Given π_k, z_{dk} is sampled independently from a Bernoulli distribution, that is $\forall d,\ z_{dk} \sim \text{Bernoulli}(\pi_k)$. For developing a variational method to learn iEFH, the parameters $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots, \pi_K\}$ are generated by a stick-breaking process [3]: $\pi_1 = v_1$, and $\pi_k = v_k \pi_{k-1} = \prod_{i=1}^k v_i,$, where $v_i \sim \text{Beta}(\alpha, 1)$. Then, the prior distribution is $p(Z, \mathbf{v}) = p(\mathbf{v}) p(Z|\mathbf{v}) = p(\mathbf{v}) \prod_d p(\mathbf{z}_d|\mathbf{v})$ and $p(\mathbf{z}_d|\mathbf{v}) = \prod_k p(z_{dk}|\mathbf{v}) = \prod_k \text{Bernoulli}(\pi_k)$.

Optimization with Contrastive Divergence: To avoid the intractable normalization factor, we develop an efficient contrastive divergence inference method [4][1] and use the following objective to approximate the negative log-likelihood

$$\mathcal{L}(\Theta, q_0, q_1) \triangleq \text{KL}(q_0(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\}) \| p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\})) - \text{KL}(q_1(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\}) \| p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\}))$$

where $\Theta = (\mathbf{W}, \alpha, \beta)$ is the model parameters, $\text{KL}(q, p)$ is the KL-divergence of variational distribution q (q_0 or q_1) and model distribution p. q_0 is defined with $\mathbf{x}$ and $\mathbf{z}$ clamped to the observed values and q_1 is defined with all the variables free. By using IBP prior, the joint distribution $p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\})$ of the single-view iEFH model is $p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{z}_d\}) \propto p(\mathbf{v}) \prod_d p(\mathbf{z}_d|\mathbf{v}) \exp\{\boldsymbol{\alpha}^\top \mathbf{x}_d - \frac{1}{2}\mathbf{x}_d^\top \mathbf{x}_d + \boldsymbol{\beta}^\top \mathbf{z}_d + \mathbf{x}_d^\top \mathbf{W} \mathbf{z}_d\}$, we update each variational distribution under the mean field assumption. Each of the factored distribution has the



(a) (b) (c)

Figure 1: (a) iEFH using the stick-breaking construction of IBP prior; classification accuracy on (b) TRECVID 2003, (c) 13-animal Flickr image datasets.

following form: $q(x_{dn}) = Normal(x_{dn};\ \mathbb{E}_q[x_{dn}])$; $q(z_{dk}) = Bernoulli(z_{dk}; \nu_{dk})$; $q(v_k) = Beta(v_k; \gamma_{k1}, \gamma_{k2})$. For the variational parameters ν and $\boldsymbol{\gamma}$, we need to evaluate the terms:

$$\mathbb{E}_q[\log p(\mathbf{v})] = \sum_k \mathbb{E}_q[\log p(v_k)] = (\alpha - 1) \sum_k \mathbb{E}_q[\log v_k] + c$$

$$\mathbb{E}_q[\log p(\mathbf{z}_d|\mathbf{v})] = \sum_k \mathbb{E}_q[\log p(z_{dk}|\mathbf{v})] = \sum_k \Big\{ \nu_{dk} \sum_{j=1}^{k} \mathbb{E}_q[\log v_j] + (1 - \nu_{dk}) \mathbb{E}_q[\log(1 - \prod_{j=1}^{k} v_j)] \Big\},$$

where $\mathbb{E}_q[\log v_k] = \psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2})$ and ψ is the digamma function. $c = K(\log\Gamma(\alpha+1) - \log\Gamma(\alpha)) = K \log\alpha$ is a constant. Then the mean-field update equation for $\mathbb{E}_q[\mathbf{Z}] = \nu$ is defined as: $\nu_{dk} = 1/\{1 + \exp\{\tau_1 - \tau_2 - (\beta_k + \mathbb{E}_q[\mathbf{x}_d]^\top \mathbf{W}_{.k})\}\}$, where $\tau_1 = \mathbb{E}_q[\log(1 - \prod_{j=1}^{k} v_j)]$, and $\tau_2 = \sum_{j=1}^{k} \mathbb{E}_q[\log v_j]]$. We use the multinomial bound [2] when $k > 1$ to compute τ_1. The update equation for $\boldsymbol{\gamma}$ has the same form as in [2].

After we achieve q_0 and q_1 with contrastive divergence, parameter learning can be done by optimizing the approximate objective function $\mathcal{L}(\Theta, q_0, q_1)$ using gradient descent. Let $\Delta\mathbb{E}[\cdot] = \mathbb{E}_{q_1}[\cdot] - \mathbb{E}_{q_0}[\cdot]$, then the gradients of model parameters Θ are:

$$\nabla_{\alpha_n}\mathcal{L} = \sum_d \Delta\mathbb{E}[x_{dn}],\ \ \nabla_{\beta_k}\mathcal{L} = \sum_d \Delta\mathbb{E}[z_{dk}],\ \ \nabla_{\mathbf{W}_{nk}}\mathcal{L} = \sum_d \Delta\mathbb{E}[x_{dn} z_{dk}].$$

Note that the sparseness of the model depends on the non-zero terms of the latent features Eq[z], which is influenced by the IBP prior. Due to space limit, derivations for multi-view iEFH and max-margin iEFH (iMMH) are deferred to longer version of the paper.

3 Experiments and Discussions

We have presented an infinite exponential family Harmonium (iEFH) model. Now, we provide several experiments to evaluate the proposed model on two real-valued dataset consisting of multi-view features, including TRECVID 2003 and 13-animal Flickr image [1]. We compare the iMMH, iEFH with the supervised harmonium (MMH) [1] on text feature, color feature and multi-view features respectively. Classification accuracy is showed in Fig. 1(b) and Fig. 1(c). iMMH performs consistently comparable with (or even better than) MMH whose number of latent features is determined by the model selection procedure, and iMMH consistently outperforms iEFH which uses unsupervised maximum likelihood estimation. We also analyze the sensitivity of the infinite Harmonium models with respect to the hyper-parameters (e.g., α) in the longer version of the paper.

References

[1] Ning Chen, Jun Zhu, and Eric P. Xing. Predictive subspace learning for multi-view data: A large margin approach. In *NIPS*, 2010.

[2] Finale Doshi-Velez, Kurt T. Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the indian buffet process. In *AISTATS*, 2009.

[3] Yee Whye Teh, Dilan Gorur, and Zoubin Ghahramani. Stick-breaking construction of the indian buffet process. In *AISTATS*, 2007.

[4] Max Welling and Geoffrey E. Hinton. A new learning algorithm for mean field boltzmann machines. In *ICANN*, 2001.