A Discriminative Nonparametric Bayesian Model: Infinite Hidden Conditional Random Fields

Konstantinos Bousmalis
Imperial College London
k.bousmalis@imperial.ac.uk

Louis–Philippe Morency[*]
University of Southern California
morency@ict.usc.edu

Stefanos Zafeiriou[*]
Imperial College London
s.zafeiriou@imperial.ac.uk

Maja Pantic
Imperial College London, U.Twente
m.pantic@imperial.ac.uk

1 Introduction

Nonparametric methods have been successfully applied to many existing graphical models with latent variables [3, 2, 7, 4]. In contrast to all previous work, the infinite Hidden Conditional Random Fields (iHCRF), introduced in this work, is the first, to our knowledge, discriminative bayesian nonparametric model.

Finite Hidden Conditional Random Fields (HCRFs) [5] are discriminative models that learn the joint distribution of a class label and a sequence of latent variables conditioned on a given observation sequence, with dependencies among latent variables expressed by an undirected graph. A limitation of the finite HCRFs is that finding the optimal number of hidden states for a given classification problem is not always intuitive, and involves cross–validation, that can be very computationally expensive. This limitation motivated our nonparametric HCRF model that automatically learns the optimal number of hidden states given a specific dataset. This is achieved by using Hierarchical Dirichlet Processes (HDPs) to allow for an infinite number of hidden states for the HCRF. The reader is encouraged to look at [6] for a complete description of Hierarchical Dirichlet Processes [6].

2 Infinite Hidden Conditional Random Fields (iHCRFs)

We represent T observations as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$. Each observation at time $t \in \{1, \dots, T\}$ is represented by a feature vector $\mathbf{f}(\mathbf{X}, t) \in \Re^d$, where d is the number of features. $\mathbf{f}_t$, for brevity, is a vector that can include any features of the observation sequence. One of the main representational power of iHCRFs is that the latent variables can depend on arbitrary features of the observation sequence. We wish to learn a mapping between observation sequence $\mathbf{X}$ and class label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of available labels. The iHCRF does so by estimating the conditional joint distribution over a sequence of latent variables $\mathbf{s} = [s_1, s_2, \dots, s_T]$ and a label y, given $\mathbf{X}$. The iHCRF allows its latent variables to select from an infinite number of hidden states $\mathcal{H} = \{h_1, h_2, \dots, h_\infty\}$ and learn an appropriate model configuration automatically. Being discriminative, the iHCRF models the conditional probability of a class label given an observation sequence by: $P(y \mid \mathbf{X}; \boldsymbol{\theta}) = \sum_{\mathbf{s}} P(y, \mathbf{s} \mid \mathbf{X}; \boldsymbol{\theta}) = \frac{\sum_{\mathbf{s}} \Psi(y,\mathbf{s},\mathbf{X};\boldsymbol{\theta})}{\sum_{y' \in \mathcal{Y},\mathbf{s}} \Psi(y',\mathbf{s},\mathbf{X};\boldsymbol{\theta})}$.The potential function $\Psi(y, \mathbf{s}, \mathbf{X}; \boldsymbol{\theta}) \in \Re$ is parameterized by $\boldsymbol{\theta}$, which measures the compatibility between a label, a sequence of observations and a configuration of the hidden states. In this work, the graph of our model is a chain where each node corresponds to a s_t. Our parameter vector $\boldsymbol{\theta}$ is made up of three components: $\boldsymbol{\theta} = [\boldsymbol{\theta}^{\boldsymbol{x}T}\ \boldsymbol{\theta}^{\boldsymbol{y}T}\ \boldsymbol{\theta}^{\boldsymbol{e}T}]^T$. $\boldsymbol{\theta}^{\boldsymbol{x}}$ models the compatibility between features and hidden states. $\boldsymbol{\theta}^{\boldsymbol{y}}$ models the compatibility of the hidden states and labels . $\boldsymbol{\theta}^{\boldsymbol{e}}$ models the compatibility of state transitions and labels. It is the equivalent to the transition matrix in a iHMM, but an important difference is that an HCRF model keeps a matrix of "transition" weights for each label. We define potential functions for each of these relationships, and our Ψ as their product along the chain:

$$\Psi(y, \mathbf{s}, \mathbf{X}; \boldsymbol{\theta}) = \Psi^x(\mathbf{s}, \mathbf{X}; \boldsymbol{\theta}^x)\Psi^y(y, \mathbf{s}; \boldsymbol{\theta}^y)\Psi^e(y, \mathbf{s}; \boldsymbol{\theta}^e) \tag{1}$$

[*] Stefanos Zafeiriou and Louis–Philippe Morency had an equal contribution to this work

$$\Psi^x(\mathbf{s}, \mathbf{X}; \boldsymbol{\theta}^x) = \prod_{t=1}^{T} \exp\{\mathbf{f}_t \cdot \boldsymbol{\theta}^x[s_t]\}, \quad \Psi^y(y, \mathbf{s}; \boldsymbol{\theta}^y) = \prod_{t=1}^{T} \exp\{\theta^y[y, s_t]\}, \quad \Psi^e(y, \mathbf{s}; \boldsymbol{\theta}^e) = \prod_{t=2}^{T} \exp\{\theta^e[y, s_\lambda, s_\kappa]\}$$

where $\mathbf{a} \cdot \mathbf{b}$ denotes the dot product between vectors $\mathbf{a}$ and $\mathbf{b}$. We use the notation $\boldsymbol{\theta^x}[h_i]$ to refer to the weights that corresponds to state h_i. Similarly, $\theta^y[y, h_i]$ refers to parameters that correspond to class y and state h_i, and $\theta^e[y, h_i, h']$ measures the compatibility of a transition from h_i to h' and y. In our iHCRF $\boldsymbol{\theta}^x$, $\boldsymbol{\theta}^y$, and $\boldsymbol{\theta}^e$ are obtained as a function of the mixing proportions produced from three Hierarchical Dirichlet Processes [6], a separate process for each of our potential functions in (1): HDP^x, HDP^y, HDP^e. The choice to use HDPs and not separate DPs was in line with previous work (e.g. [3]) as for each set of the iHCRF parameter sets we want to introduce intraset dependencies. These dependencies should be different for each of the three components of $\boldsymbol{\theta}$, which was also the reason for our choosing three distinct HDPs with different hyperparameters and counts. The number of visited states represented in an iHCRF, K, is nevertheless the same for all three HDPs. Our iHCRF is then parameterized only by the 6 hyperparameters —$\{\alpha_0^x, \gamma^x, \alpha_0^y, \gamma^y, \alpha_0^e, \gamma^e\}$:

$$P(y, \mathbf{s} \mid \mathbf{X}; \alpha_0^x, \gamma^x, \alpha_0^y, \gamma^y, \alpha_0^e, \gamma^e) \;\propto\; \Psi^x(\mathbf{s}, \mathbf{X}; \boldsymbol{\theta}^x)\Psi^y(y, \mathbf{s}; \boldsymbol{\theta}^y)\Psi^e(y, \mathbf{s}; \boldsymbol{\theta}^e) \tag{2}$$

$$\boldsymbol{\beta^x} \sim GEM(\gamma^x), \qquad \boldsymbol{\pi_j^x} \mid \boldsymbol{\beta^x} \;\sim\; DP(\alpha_0^x, \boldsymbol{\beta^x}), \qquad \boldsymbol{\theta_j^x} = K\boldsymbol{\pi_j^x} - 1 \tag{3}$$

$$\boldsymbol{\beta^y} \sim GEM(\gamma^y), \qquad \boldsymbol{\pi_j^y} \mid \boldsymbol{\beta^y} \;\sim\; DP(\alpha_0^y, \boldsymbol{\beta^y}), \qquad \boldsymbol{\theta_j^y} = K\boldsymbol{\pi_j^y} - 1 \tag{4}$$

$$\boldsymbol{\beta^e} \sim GEM(\gamma^e), \qquad \boldsymbol{\pi_j^e} \mid \boldsymbol{\beta^e} \;\sim\; DP(\alpha_0^e, \boldsymbol{\beta^e}), \qquad \boldsymbol{\theta_j^e} = K\boldsymbol{\pi_j^e} - 1 \tag{5}$$

The functions for $\boldsymbol{\theta}^x, \boldsymbol{\theta}^y, \boldsymbol{\theta}^e$ were chosen to allow for negative weights.

For hyperparameter learning and inference, we adapted the Beam Sampling algorithm, an MCMC sampling method that has successfully been used to sample whole trajectories for the iHMM [3]. Since we have adopted in this work a chain structure for our model, the Beam Sampler can easily be adapted for the iHCRF, by forward filtering and backward sampling on an undirected chain.

3 Experiments

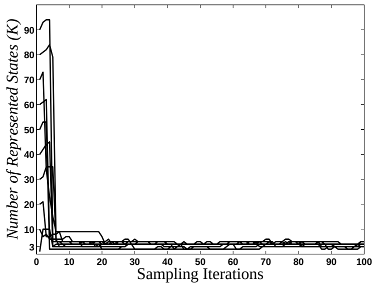

Figure 1: Convergence analysis of iHCRF number of represented states K.

We evaluated the performance of the proposed iHCRF and the adapted Beam Sampler on the recognition of agreement and disagreement episodes in video sequences [1]. We conducted iHCRF experiments with initial values for $K = \{1, 10, 20, \cdots, 100\}$ spliting our dataset equally between training and testing sets. As can be seen in figure 1, our model was able to converge within 20 iterations, in all cases, to a K ranging from 3 to 5 states. The iHCRF model with initial $K = 10$ converged to 3 states and achieved 69.23% total accuracy, recognizing 76.92% of agreement and 61.54% of disagreement test cases. We trained and tested finite HCRFs, with hidden states 2, 3, 5, 7 and 9, and 11 different random initializations, using the same datasplits. The best total accuracy was 61% and it was shown that converging to a model with 3 states was the appropriate choice, same as our iHCRF model.

References

[1] K. Bousmalis, L.-P. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *IEEE FG 2011*.

[2] E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. PhD thesis, MIT, July 2009.

[3] J. Van Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden markov model. In *ICML 2008*.

[4] P. Orbanz and J. Buhmann. Nonparametric bayes image segmentation. *Int'l Journal in Computer Vision*, 77:25–45, 2008.

[5] A. Quattoni, S. Wang, L.P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE PAMI*, pages 1848–1852, 2007.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and Blei D.M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[7] J. Van Gael, Y.W. Teh, and Z. Ghahramani. The infinite factorial hidden markov model. *Advances in Neural Information Processing Systems*, 21:1697–1704, 2009.