Variational Nonparametric Bayesian Hidden Markov Model

Nan Ding, Zhijian Ou

ozj@tsinghua.edu.cn

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

2010 ICASSP Dallas · IEEE · 2010 IEEE International Conference on Acoustics, Speech and Signal Processing · March 14 - 19, 2010 • Sheraton Dallas Hotel • Dallas, Texas, U.S.A.

Objective: discover the structure of the HMM state space.

Method: propose a nonparametric Bayesian HMM (NBHMM) based on Dirichlet Process

Advantage: theoretically sound, efficient computation with variational inference

Differences from other existing nonparametric Bayesian HMM

iHMM: Beal, Ghahramani, Rasmussen, "The infinite hidden Markov model," NIPS 2002.

HDP-HMM: Teh, Jordan, Beal, Blei, "Hierarchical Dirichlet processes," JASA 2006.

1	iHMM and HDP-HMM employ sampling based inference.	We apply the efficient variational inference for the NBHMM.
2	iHMM deals only with discrete observations.	NBHMM supports continuous observations via (infinite) Gaussian mixtures.
3	The transition distribution in iHMM and HDP-HMM is generated from HDP	In the NBHMM, directly created from a stickbreaking construction, simpler

Variational Inference on NBHMM

Basic Idea: minimize the Kullback-Leibler distance KL(q|p)

p is the true distribution

q is the approximate distribution

$$p(\mathbf{s},\mathbf{h},\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma|\mathbf{o})$$
$$\approx q(\mathbf{s},\mathbf{h})q(\boldsymbol{\pi}')q(\mathbf{A}')q(\mathbf{C}')q(\boldsymbol{\mu},\Sigma^{-1})$$
$$= q(s_1)\prod_{t=2}^{T} q(s_t|s_{t-1})\prod_{t=1}^{T} q(h_t|s_t)$$
$$\cdot\prod_{i=1}^{L} q(\pi'_i)\prod_{j=1}^{L}\prod_{i=1}^{L} q(a'_{ji})\prod_{j=1}^{L}\prod_{k=1}^{L} q(c'_{jk})$$
$$\cdot\prod_{j=1}^{L}\prod_{k=1}^{L}\prod_{d=1}^{D} q(\mu_{jkd}|\Sigma^{-1}_{jkd})q(\Sigma^{-1}_{jkd})$$

where,

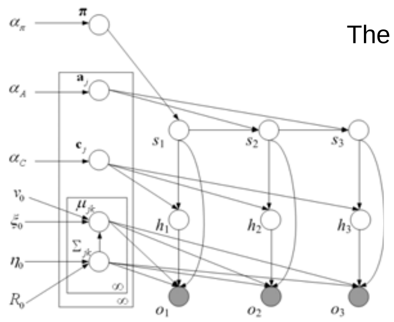
$$q(\pi'_i) = Beta(\tau_{1(\pi'_i)}, \tau_{2(\pi'_i)})$$
$$q(a'_{ji}) = Beta(\tau_{1(a'_{ji})}, \tau_{2(a'_{ji})})$$
$$q(c'_{jk}) = Beta(\tau_{1(c'_{jk})}, \tau_{2(c'_{jk})})$$
$$q(\mu_{jkd}|\Sigma^{-1}_{jkd}) = \mathcal{N}(\tilde{v}_{jkd}, \tilde{\xi}^{-1}_{jkd}\Sigma_{jkd})$$
$$q(\Sigma^{-1}_{jkd}) = Gamma(\tilde{\eta}_{jkd}, \tilde{R}_{jkd})$$

Two variational assumptions:

- Assume $(\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma)$ and $(\mathbf{s},\mathbf{h})$ are mutually independent.
- We only compute the posterior probabilities for *L* states of the infinite large state-space. Only the states corresponding to "large" posteriors are effective in explaining the observed data.

Nonparametric Bayesian HMM (NBHMM)

Graphical model of the Bayesian HMM



The joint distribution is

$$p(\mathbf{s},\mathbf{h},\mathbf{o},\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma^{-1}) = p(\mathbf{s},\mathbf{h},\mathbf{o}|\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma)p(\boldsymbol{\pi})p(\mathbf{A})p(\mathbf{C})p(\boldsymbol{\mu}|\Sigma^{-1})p(\Sigma^{-1}) \quad (1)$$

where $\mathbf{s} = (s_t)_{t=1}^T$, $\mathbf{h} = (h_t)_{t=1}^T$, $\mathbf{o} = (o_t)_{t=1}^T$.

$$p(\mathbf{s},\mathbf{h},\mathbf{o}|\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma^{-1}) = p(s_1|\boldsymbol{\pi})\prod_{t=2}^{T} p(s_t|s_{t-1},\mathbf{A})\prod_{t=1}^{T} p(h_t|s_t,\mathbf{C})p(o_t|\mu_{s_t,h_t},\Sigma_{s_t,h_t}) \quad (2)$$

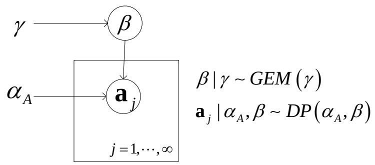
Gaussian-Gamma prior on the Gaussian components :

$$p(\mu_{jkd}|\Sigma^{-1}_{jkd}) = \mathcal{N}(v_0, \xi_0^{-1}\Sigma_{jkd})$$
$$p(\Sigma^{-1}_{jkd}) = Gamma(\eta_0, R_0)$$

A stickbreaking construction of the Dirichlet Process prior for the transition matrix :

$$p(\pi'_i) = Beta(1,\alpha_\pi) \qquad \pi_i = \pi'_i\prod_{n=1}^{i-1}(1-\pi'_n)$$
$$p(a'_{ji}) = Beta(1,\alpha_A) \qquad a_{ji} = a'_{ji}\prod_{n=1}^{i-1}(1-a'_{jn})$$
$$p(c'_{jk}) = Beta(1,\alpha_C) \qquad c_{jk} = c'_{jk}\prod_{l=1}^{k-1}(1-c'_{jl})$$

- The stickbreaking distribution over $\mathbf{a}_j$: $\mathbf{a}_j \sim GEM(\alpha_A)$
- HDP

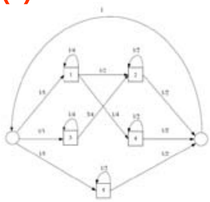


$\beta|\gamma \sim GEM(\gamma)$

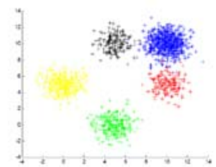
$\mathbf{a}_j|\alpha_A,\beta \sim DP(\alpha_A,\beta)$

Experiment Results

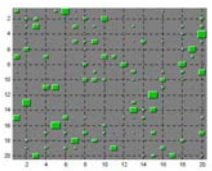
(1) NBHMM vs. Classical HMM



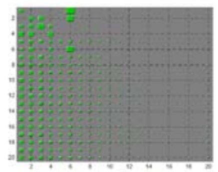
(a) Synthetic Markov machine.



(b) Synthetic observations

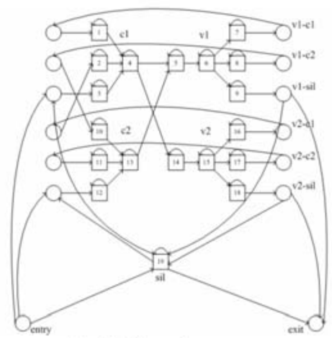


(c) Hinton graph for classical HMM

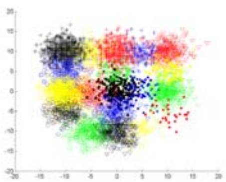


(d) Hinton graph for NBHMM

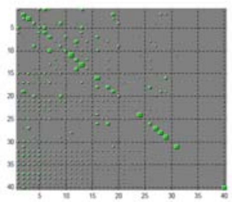
(2) Triphone model



(a) A "triphone" structure



(b) Synthetic observations



(c) Hinton graph for NBHMM

(3) Chinese isolated syllable recognition

- **1254 syllables, whole-syllable HMMs**
- **(14 MFCCs+ E)*3 = 45-dim feature**
- **50 males data, leave-one-out test**
- **classic HMM: 6-state 73.4%, 16-state 80.1%**
- **NBHMM: discover 14-18 effective states for the syllables, 78.9% accuracy**