Max-Margin Infinite Hidden Markov Models

Aonan Zhang ZAN12@TSINGHUA.EDU.CN
Jun Zhu DCSZJ@TSINGHUA.EDU.CN
Bo Zhang DCSZB@TSINGHUA.EDU.CN
Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys., Tsinghua University, China

Abstract

Infinite hidden Markov models (iHMMs) are nonparametric Bayesian extensions of hidden Markov models (HMMs) with an infinite number of states. Though flexible in describing sequential data, the generative formulation of iHMMs could limit their discriminative ability in sequential prediction tasks. Our paper introduces max-margin infinite HMMs (M2iHMMs), new infinite HMMs that explore the max-margin principle for discriminative learning. By using the theory of Gibbs classifiers and data augmentation, we develop efficient beam sampling algorithms without making restricting mean-field assumptions or truncated approximation. For single variate classification, M2iHMMs reduce to a new formulation of DP mixtures of max-margin machines. Empirical results on synthetic and real data sets show that our methods obtain superior performance than other competitors in both single variate classification and sequential prediction tasks.

1. Introduction

Hidden Markov models (HMMs) (Rabiner, 1989) are one of the most well-known methods for modeling sequential data, such as speech and videos, by using a Markov chain to capture the dynamic dependencies among data. Statistics of the latent states can be inferred by either a maximum likelihood treatment or Bayesian formulation with efficient forward-backward algorithms. Recently, by using the theory of (hierarchical) Dirichlet processes (DPs) (Ferguson, 1973; Antoniak, 1974), extensions have been made to derive infinite HMMs (iHMMs) (Beal et al., 2001), which allow the models to have an unbounded number of hidden states. The posterior distribution of iHMMs can be inferred with a Gibbs sampler (Beal et al., 2001; Teh et al., 2006) or a more efficient beam sampler (Gael et al., 2008).

Although HMMs and iHMMs are flexible in capturing interdependencies in sequential data, their generative formulations could limit the discriminative ability in sequential prediction tasks, e.g., speech recognition and object tracking in videos. Successful attempts have been made to perform discriminative training for HMMs, including max-margin Gaussian mixture HMMs for speech recognition (Sha & Saul, 2006) and large-margin Markov models for structured output prediction (Altun et al., 2004; Taskar et al., 2003). But these approaches learn a single large-margin model, which can be insufficient to capture the underlying structures (e.g., sequential clusters) when data have complex dynamics. Furthermore, the non-Bayesian formulations make these approaches not obviously generalizable to nonparametric Bayesian iHMMs.

To discover underlying descriptive patterns (e.g., clusters) and/or improve efficiency, much progress has been made in single variate classification. For example, mixture-of-experts (Collobert et al., 2002; Fu et al., 2010) models have been developed to partition the observation space into subregions and learn a SVM classifier within each subregion. Recently, inspired by the success of DP mixtures of generalized linear models (Shahbaba & Neal, 2009; Hannah et al., 2010), Zhu et al. (2011) proposed infinite SVMs (iSVMs), DP mixtures of large-margin machines, which inherit the advantages of Bayesian nonparametrics to resolve the unknown number of components and the large-margin principle to learn discriminative classifiers.

Though not considering interdependencies among data, iSVMs offer a promising direction to incorporate a potentially infinite mixture-of-experts in HMMs for sequential prediction tasks. Recent work on infinite Markov-switching maximum entropy discrimination machines (iM2EDMs) (Chatzis, 2013) extend iSVMs to capture the sequential dependencies by connecting the latent cluster assignment variables via a Markov chain. iM2EDMs follow the suggestions of iSVMs and build large-margin models by using expected/averaging discriminant functions. Since it is intractable to deal with an infinite number of la-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

tent states as well as the non-smooth hinge loss function, iM2EDMs adopt truncated variational inference with a factorized mean-field assumption, which could be a poor approximation to the true posterior.

This paper presents max-margin infinite hidden Markov models (M2iHMMs), new max-margin extensions of the nonparametric Bayesian iHMMs for sequential prediction, which admit efficient Markov chain Monte Carlo (MCMC) inference algorithms without making truncated approximation or mean-field assumptions. Technically, we build M2iHMMs as regularized Bayesian extensions to iHMMs by using the ideas of Gibbs classifiers (Langford & Shawe-Taylor, 2003; Germain et al., 2009; Zhu et al., 2014) to derive a max-margin posterior regularization term, which is a good surrogate (in fact, an upper bound) of the training error. By exploring data augmentation techniques (Tanner & Wong, 1987; Polson & Scott, 2011), we are able to develop efficient MCMC algorithms with a beam sampler to efficiently deal with the Markov chain dynamics. For the special case of single variate classification, M2iHMMs reduce to Gibbs infinite SVMs (GiSVMs), new formulations of DP mixtures of large-margin machines, with a truncation-free and assumption-free Gibbs sampling algorithm. Moreover, the expected hinge loss in GiSVMs is an upper bound of the hinge loss of an expected/averaging classifier in iSVMs. Finally, experimental results on both synthetic and real data sets demonstrate superior performance of our approaches in both single variate classification and sequential prediction tasks, compared to various competitors.

2. Infinite Hidden Markov Models

We start with a brief overview of HMMs and infinite HMMs. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ denote an observed sequence of length T, and each single observation $\mathbf{x}_t \in \mathbb{R}^M$ is a feature vector. An HMM model defines a joint distribution over $\mathbf{X}$ by invoking another sequence of hidden discrete state variables $\mathbf{Z} = \{z_0 = 1, z_1, z_2, \cdots, z_T\}$[1], and each z_t takes values from a finite set with K values, e.g., $\{1, \cdots, K\}$. For the common first-order Markov models, the basic assumption is that given z_t, z_i is independent of z_j for all $i < t < j$. This Markov dynamics is formally characterized by a transition probability distribution

$$p(z_t = j | z_{t-1} = i, \boldsymbol{\pi}) = \pi_{ij}, \quad t = 1, 2, \cdots, T, \tag{1}$$

where $\boldsymbol{\pi}$ is a $K \times K$ transition matrix. Given the hidden states, the observations $\mathbf{X}$ are modeled by an emission distribution, e.g., a normal distribution for real-valued inputs

$$p(\mathbf{x}_t | z_t, \boldsymbol{\gamma}) = \frac{1}{(2\pi|\Sigma_{z_t}|)^{M/2}} \exp\big(-D_{z_t}(\mathbf{x}_t|\boldsymbol{\gamma})\big), \tag{2}$$

[1] z_0 is an initial state.

where $D_{z_t}(\mathbf{x}_t|\boldsymbol{\gamma}) = \frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{z_t})^\top \boldsymbol{\Sigma}_{z_t}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{z_t})$, and $\boldsymbol{\gamma}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are the mean and covariance parameters of the likelihood model in component k. Then the joint probability distribution induced by HMM is

$$p(\mathbf{X}, \mathbf{Z}) = \prod_{t=1}^{T} p(z_t|z_{t-1}) p(\mathbf{x}_t|z_t). \tag{3}$$

The dependencies among observations are captured via the Markov chain on $\mathbf{Z}$, whose statistics can then be inferred by an efficient forward-backward message passing scheme within an EM algorithm for maximum likelihood estimation (Rabiner, 1989). To make inference and prediction more robust, Bayesian HMMs have been examined by introducing priors $p_0(\boldsymbol{\pi}|\boldsymbol{\beta})$ and $p_0(\boldsymbol{\gamma}|\Gamma)$ (Scott, 2002).

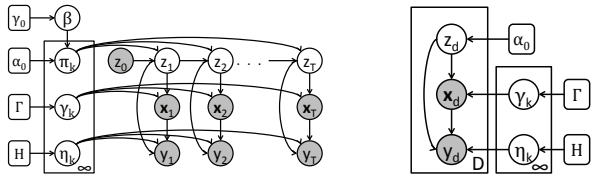
One limitation of HMMs is that the number of hidden states needs to be pre-specified. Infinite HMMs (iHMMs) were proposed (Beal et al., 2001), in which a hierarchical Dirichlet process (HDP) (Teh et al., 2006) is used as a nonparametric prior for the transition matrix $\boldsymbol{\pi}$ to allow a countably infinite number of coupling rows[2]. Formally, HDP combines an infinite number of Dirichlet processes (DP) where each row of the transition matrix $\boldsymbol{\pi}_k$ is drawn from a Dirichlet process with a shared base measure $\boldsymbol{\beta}$, which follows a stick-breaking construction (Pitman, 2002), i.e.,

$$\boldsymbol{\pi}_k | \boldsymbol{\beta} \sim \mathcal{DP}(\alpha_0, \boldsymbol{\beta}), \ \boldsymbol{\beta} \sim \text{GEM}(\gamma_0). \tag{4}$$

The hyper-parameters α_0 and γ_0 can be either set a priori or inferred via a fully Bayesian treatment by introducing hyper-priors (e.g., Gamma priors). For posterior inference, we can use the theory of HDP to perform Gibbs sampling (Teh et al., 2006), or we can apply a more efficient beam sampler (Gael et al., 2008) by using an augmented representation of the Markov chain.

Though iHMMs are flexible in modeling sequential data with an unbounded number of hidden states, the generative nature could make them less than sufficient in learning discriminative models for sequential prediction tasks. To improve the discriminative ability, successful attempts have been made on developing discriminative HMMs, e.g., by exploring max-margin learning in an optimization framework for finding a single large-margin model (Sha & Saul, 2006; Altun et al., 2004; Taskar et al., 2003). But these approaches are not easy to generalize to the nonparametric Bayesian iHMMs. The recent work (Chatzis, 2013) provides one solution, but as stated above, its inference algorithm relies on truncation and strict mean-field assumptions. Below, we present a new formulation of max-margin infinite HMMs and provide an efficient MCMC algorithm without truncation or mean-field assumptions.

[2] The rows are coupled to allow dependencies over transitions, which is essential to provide non-trivial solutions through inference when the number of hidden states goes to infinity.



(a) (b)

Figure 1. Graphical model representation of (a) M2iHMM and (b) GiSVM. Note that M2iHMM reduces to GiSVM when we do not assume the sequential dependencies among the latent variables for single variate classification.

3. Max-Margin Infinite HMMs

A max-margin iHMM is a nonparametric Bayesian HMM for sequential prediction (see Fig. 1(a)), in which the max-margin principle is explored to regularize posterior inference for better discriminative ability. For the ease of understanding, we start with single variate classification (see Fig. 1(b)), where observations are modeled separately by using DP mixtures. We then extend max-margin DP mixtures to model sequential data by building a Markov chain process with the HDP theory.

3.1. Gibbs iSVM for Single Variate Classification

For single variate classification, the training set consists of D i.i.d samples $(\mathbf{x}_d, y_d)$, where $\mathbf{x}_d \in \mathbb{R}^M$ is an input feature vector and $y_d \in \mathcal{Y} = \{1, \cdots, L\}$ is a discrete response variable for multi-way classification. Gibbs iSVM is a DP mixture model for describing both input features and response variables, as illustrated in Fig. 1(b). It consists of two parts — a DP mixture for input features and a Gibbs classifier for response variables, as explained below.

3.1.1. DP MIXTURES

Let z_d denote the component assignment for data point d. A DP mixture consists of a likelihood model $p(\mathbf{x}_d|z_d, \boldsymbol{\gamma})$ similar as Eq. (2) for describing the observed data in each cluster, a Chinese Restaurant Process (CRP) prior (Pitman, 1995) over the latent variables $\mathbf{Z}$, and some priors over parameters $\boldsymbol{\gamma}$. Given a set of data $\mathbf{X}$, we can apply Bayes' rule to infer the posterior distribution, $p(\mathbf{Z}, \boldsymbol{\gamma}|\mathbf{X})$. From the variational point of view, the posterior distribution via Bayes' rule is equivalent to the solution of a convex optimization problem

$$\min_{q(\mathbf{Z},\boldsymbol{\gamma})\in\mathcal{P}} \mathrm{KL}(q(\mathbf{Z},\boldsymbol{\gamma})\|p_0(\mathbf{Z},\boldsymbol{\gamma})) - \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z},\boldsymbol{\gamma})], \quad (5)$$

where $\mathcal{P}$ is a probability simplex. We should note that the variational re-formulation doesn't reduce the complexity of doing posterior inference, and we still need to perform variational approximation or Monte Carlo methods in practice. But the significance is that it provides a nice starting point to augment DP mixtures for the discriminative max-margin learning, as detailed below.

3.1.2. REGULARIZED DP MIXTURES

To augment the DP mixtures for prediction tasks, we define a classifier over y within each cluster. Previous work has either built a likelihood model (e.g., logistic regression) for y (Shahbaba & Neal, 2009) or a large-margin classifier with an expected discriminant function (Zhu et al., 2011) to account for the uncertainty of $\mathbf{Z}$. We present a new formulation and discuss its relations to iSVM.

Let $\boldsymbol{\eta}_k$ be the classifier weights in cluster k. If we have known the component assignment z_d and the classifier $\boldsymbol{\eta}_{z_d}$, we can define some prediction rule to classify data d. We consider the simple linear discriminant function

$$f(y, \mathbf{x}_d; z_d, \boldsymbol{\eta}) = \boldsymbol{\eta}_{z_d}^\top g(y, \mathbf{x}_d) = \sum_{k=1}^{\infty} \delta_{z_d,k} \boldsymbol{\eta}_k^\top g(y, \mathbf{x}_d), \quad (6)$$

where $g(y, \mathbf{x}_d)$ is a long vector consisting of L subvectors with the y-th being $\mathbf{x}_d$ and all others being zero, and make predictions using the rule $\hat{y}_d = \arg\max_y f(y, \mathbf{x}_d; z_d, \boldsymbol{\eta})$. Following the approach by Crammer & Singer (2001), we define the multi-class hinge loss

$$\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta}) = \sum_d \max_y (\Delta_d^y - \Delta f(y, y_d, \mathbf{x}_d; z_d, \boldsymbol{\eta})), \quad (7)$$

where Δ_d^y equals to 0 if $y = y_d$ and ℓ (≥ 1) otherwise; ℓ is the cost of making a wrong prediction; and $\Delta f(y, y_d, \mathbf{x}_d; z_d, \boldsymbol{\eta}) = f(y_d, \mathbf{x}_d; z_d, \boldsymbol{\eta}) - f(y, \mathbf{x}_d; z_d, \boldsymbol{\eta})$ is the margin favored by the ground truth y_d over any other label y. This hinge loss is convex with respect to $\boldsymbol{\eta}$ and also an upper bound of the training cost, $\sum_d \ell(1 - \delta_{y_d,\hat{y}_d})$. To account for the uncertainty of $\mathbf{Z}$ and $\boldsymbol{\eta}$, we adopt the approach of Gibbs classifiers (Germain et al., 2009; Zhu et al., 2013; 2014) and take expectation over the target posterior $q(\mathbf{Z}, \boldsymbol{\eta})$ to define the expected hinge loss

$$\mathcal{R}(q) = \mathbb{E}_q\left[\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta})\right], \quad (8)$$

which is an upper bound of the expected training cost, $\sum_d \ell \mathbb{E}_q\left[1 - \delta_{y_d,\hat{y}_d}\right]$.

With the above Gibbs classifier, we can regularize the posterior inference of DP mixtures by solving the following hybrid optimization problem

$$\min_{q(\mathbf{Z},\boldsymbol{\eta},\boldsymbol{\gamma})\in\mathcal{P}} \mathcal{L}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma})) + 2c\mathcal{R}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma})), \quad (9)$$

where $\mathcal{L}(q) = \mathrm{KL}(q\|p_0(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma})) - \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\gamma})]$ is the objective function for doing standard Bayesian inference, and c is a positive regularization constant.

Remark 1 *Unlike GiSVM, iSVM builds a max-margin DP mixture model based on averaging classifiers, which define the expected discriminant function* $f(y, \mathbf{x}_d; q) = \mathbb{E}_q[f(y, \mathbf{x}_d; z_d, \boldsymbol{\eta})]$, *and make predictions using the argmax*

rule $\hat{y}_d = \arg\max_y f(y, \mathbf{x}_d; q)$. *Let* $\Delta f(y, \mathbf{x}_d; q) = f(y_d, \mathbf{x}_d; q) - f(y, \mathbf{x}_d; q)$ *be the margin and* $\mathcal{R}' = \sum_d \max_y(\Delta_d^y - \Delta f(y, \mathbf{x}_d; q))$ *be the multi-class hinge loss of this classifier. Then, iSVM solves a hybrid optimization problem similar to (9), simply replacing* $\mathcal{R}$ *with* $\mathcal{R}'$. *In fact, we can prove that the expected hinge loss is an upper bound of the hinge loss of the expected classifier by exploring the convexity of the hinge loss, i.e.,* $\mathcal{R}(q) \geq \mathcal{R}'(q)$.

One merit for using Gibbs classifiers in GiSVM is that we can reformulate the problem with data augmentation and perform truncation-free sampling (shown in Sec. 4), which is more accurate than solving the constrained SVM subproblems in iSVM with variational approximation as in (Zhu et al., 2011).

3.2. Max-margin iHMMs for Sequential Prediction

With the above theory of max-margin DP mixtures for single variate classification, we now present the generalization of max-margin iHMMs for sequential prediction, where an instance is a pair of an observed sequence $\mathbf{X}$ and the corresponding label sequence $\mathbf{y}$. The dynamic dependency is captured by invoking another sequence of hidden states $\mathbf{Z}$, which follow a Markov chain. Similar as in iHMMs, the generative process of the latent state sequence, the observation sequence, and parameters are

$$\begin{aligned}
&\boldsymbol{\beta} \sim \text{GEM}(\gamma_0), \;\; \boldsymbol{\pi}_k|\alpha_0, \boldsymbol{\beta} \sim \mathcal{DP}(\alpha_0, \boldsymbol{\beta}), \\
&\boldsymbol{\gamma}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}|\Gamma \sim \mathcal{NIW}(\Gamma), \;\; \boldsymbol{\eta}_k|H \sim \mathcal{N}(0, H), \\
&z_t|z_{t-1}, \boldsymbol{\pi} \sim \boldsymbol{\pi}_{z_{t-1}}, \;\; \mathbf{x}_t|z_t \sim p(\mathbf{x}_t|z_t, \boldsymbol{\gamma}),
\end{aligned}$$

where Γ stands for a set of Normal-Inverse-Wishart hyper-parameters, and H is a covariance matrix, e.g., $\nu^2 I$ for isotropic Gaussian. If the cluster assignments $\mathbf{Z}$ are given, we postulate that the class labels are independently determined by the associated classifiers, and define the linear discriminant function as

$$f(\mathbf{y}, \mathbf{X}; \mathbf{Z}, \boldsymbol{\eta}) = \sum_{t=1}^{T} f(y_t, \mathbf{x}_t; z_t, \boldsymbol{\eta}), \tag{10}$$

where $f(y_t, \mathbf{x}_t; z_t, \boldsymbol{\eta})$ is the same as in (6). Then, we can make predictions using the joint argmax rule

$$\hat{\mathbf{y}}(\mathbf{z}, \boldsymbol{\eta}) = \arg\max_{\mathbf{y}} f(\mathbf{y}, \mathbf{X}; \mathbf{Z}, \boldsymbol{\eta}) \tag{11}$$

Following max-margin Markov networks (Taskar et al., 2003; Altun et al., 2004), we define the structured hinge loss for multiple sequences, each of length T, as

$$\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta}) = \sum_d \max_{\mathbf{y}} \left(\Delta_d(\mathbf{y}) - \Delta f(\mathbf{y}, \mathbf{X}_d; \mathbf{Z}_d, \boldsymbol{\eta})\right), \tag{12}$$

where $\Delta_d(\mathbf{y})$ is a cost function measuring how much $\mathbf{y}$ differs from the truth $\mathbf{y}_d^*$ for sequence d; and $\Delta f(\mathbf{y}, \mathbf{X}_d; \mathbf{Z}_d, \boldsymbol{\eta}) = f(\mathbf{y}_d^*, \mathbf{X}_d; \mathbf{Z}_d, \boldsymbol{\eta}) - f(\mathbf{y}, \mathbf{X}_d; \mathbf{Z}_d, \boldsymbol{\eta})$ is the margin favored by the ground truth $\mathbf{y}_d^*$ for sequence d. We choose the commonly used Hamming loss, that is, $\Delta_d(\mathbf{y}) = \sum_{t=1}^T \Delta_{dt}^{y_t}$, where $\Delta_{dt}^{y_t} = 1 - \delta_{y_t, y_{dt}^*}$. Due to the separability of the cost function and the discriminant function, we have the hinge loss as $\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta}) = \sum_d \sum_{t=1}^T \max_{y_t}(\Delta_{dt}^{y_t} - \Delta f(y_t, y_{dt}^*, \mathbf{x}_{dt}; z_{dt}, \boldsymbol{\eta}))$.

To resolve the uncertainty of $\mathbf{Z}$ and $\boldsymbol{\eta}$, we take the expectation and define the expected margin loss $\mathcal{R}(q(\mathbf{Z}, \boldsymbol{\eta})) = \mathbb{E}_q[\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta})]$. Then, the regularized inference problem is

$$\min_{q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\pi})} \mathcal{L}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\pi})) + 2c\mathcal{R}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi})), \tag{13}$$

an extension of (9) with a HDP process to capture the interdependencies among $\mathbf{Z}$, where $\mathcal{L}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\pi})) = \text{KL}(q \| p_0(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\pi})) - \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\gamma})]$ is the objective of Bayesian inference for HDP mixtures. We call this model M2iHMM-1 (see Fig. 1(a)). We can also use a framework similar as maximum entropy discrimination (MED) (Jaakkola et al., 1999) by omitting the likelihood model and the regularized Bayesian inference problem is

$$\min_{q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi}) \in \mathcal{P}} \mathcal{L}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi})) + 2c\mathcal{R}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi})), \tag{14}$$

where $\mathcal{L}(q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi})) = \text{KL}(q \| p_0(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi}))$ and we call this model M2iHMM-2.

4. Inference Algorithms

Now, we present truncation-free MCMC algorithms for max-margin infinite HMMs. We again start with the single variate classification for the ease of understanding.

4.1. Gibbs iSVM for Single Variate Classification

Let $\phi(y_d|z_d, \boldsymbol{\eta}) = \exp(-2c \max_y(\Delta_d^y - \Delta f(y, y_d, \mathbf{x}_d; z_d, \boldsymbol{\eta})))$ be the unnormalized likelihood of the label y_d and $\phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta}) = \prod_d \phi(y_d|z_d, \boldsymbol{\eta}_d)$. Solving problem (9), we get the normalized posterior distribution

$$q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}) = \frac{p_0(\boldsymbol{\eta}, \boldsymbol{\gamma}, \mathbf{Z}) p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\gamma}) \phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta})}{\psi(\mathbf{X}, \mathbf{y})}, \tag{15}$$

where $\psi(\mathbf{X}, \mathbf{y})$ is the normalization constant. Due to the conjugacy, we can integrate out the parameters $\boldsymbol{\gamma}$ for collapsed sampling, which may improve the mixing rate[3]. However, it would be hard to develop a MCMC algorithm for $q(\mathbf{Z}, \boldsymbol{\eta})$ directly, due to the complicated form of ϕ. Fortunately, we can develop a simple and truncation-free Gibbs sampler by exploring data augmentation techniques.

For Z: given $\boldsymbol{\eta}$, the conditional distribution is

$$q(\mathbf{Z}|\boldsymbol{\eta}) \propto p_0(\mathbf{Z}) p(\mathbf{X}|\mathbf{Z}) \phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta}), \tag{16}$$

[3] $\boldsymbol{\gamma}_k$ can be estimated using the samples assigned to cluster k once we have $\mathbf{Z}$.

where $p(\mathbf{X}|\mathbf{Z}) = \int p_0(\boldsymbol{\gamma})p(\mathbf{X}|\mathbf{Z},\boldsymbol{\gamma})d\boldsymbol{\gamma}$ is the marginal likelihood and $p_0(\mathbf{Z})$ is a CRP prior. We consider two cases to derive the conditional distribution from which a cluster assignment is drawn for data point d: 1) For the component k with the number of data points except d assigned to it $n_{-d,k} > 0$, the conditional distribution is

$$q(z_d = k|\mathbf{Z}_{-d}, \boldsymbol{\eta}) \propto n_{-d,k}\phi(y_d|z_d = k, \boldsymbol{\eta}_k) \times p(\mathbf{x}_d|Z_{-d}, \mathbf{X}^k_{-d}), \quad (17)$$

where $p(\mathbf{x}_d|\mathbf{Z}_{-d}, \mathbf{X}^k_{-d})$ is the marginal likelihood of data d being in cluster k; and 2) The probability of generating a new component k_+ is

$$q(z_d = k_+|\mathbf{Z}_{-d}) \propto \alpha_0 p(\mathbf{x}_d) \int \phi(y_d|\boldsymbol{\eta}')p_0(\boldsymbol{\eta}')\mathrm{d}\boldsymbol{\eta}', \quad (18)$$

where $p(\mathbf{x}_d) = \int p(\mathbf{x}_d|\boldsymbol{\gamma})p_0(\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\gamma}$ is the likelihood of data d, $\phi(y_d|\boldsymbol{\eta}) = \exp(-2c\max(0, \rho^{y_d}_d - f(y_d, \mathbf{x}_d; z_d, \boldsymbol{\eta}))$, and $\rho^y_d = \max_{y'\neq y}(\Delta^{y'}_d + f(y', \mathbf{x}_d; z_d, \boldsymbol{\eta}) - \Delta^y_d)$. Again, using the conjugate property, the integral $p(\mathbf{x}_d)$ can be computed in closed-form. For the second integral, we can apply importance sampling to approximate it. Then, normalizing the above terms will lead to the posterior distribution of component assignments for observation d.

For $\boldsymbol{\eta}$: given $\mathbf{Z}$, we know the number of active clusters and we can alternately sample $\boldsymbol{\eta}_k$ from the following conditional distribution by fixing other component weights

$$q(\boldsymbol{\eta}_k|\mathbf{Z}, \boldsymbol{\eta}_{-k}) \propto p_0(\boldsymbol{\eta}_k) \prod_{d:z_d=k} \phi(y_d|z_d, \boldsymbol{\eta}). \quad (19)$$

But it is still difficult to sample $\boldsymbol{\eta}_k$ from this distribution directly. Here, we develop an inner sampler to alternately sample each subvector $\boldsymbol{\eta}^y_k$ with the others fixed. Specifically, let $\zeta^y_d = \max_{y'\neq y}(\Delta^{y'}_d + f(y', \mathbf{x}_d; z_d, \boldsymbol{\eta})) - \Delta^y_d$ and $\kappa^y_d = +1$ if $y_d = y$; -1 otherwise. We can show that

$$q(\boldsymbol{\eta}^y_k|\mathbf{Z}, \boldsymbol{\eta}^{-y}_k) \propto p_0(\boldsymbol{\eta}^y_k) \prod_{d:z_d=k} \phi'(y|z_d, \boldsymbol{\eta}), \quad (20)$$

where $\phi'(y|z_d, \boldsymbol{\eta}) = \exp(-2c(\kappa^y_d\zeta^y_d - \kappa^y_d f(y, \mathbf{x}_d; z_d, \boldsymbol{\eta}))_+)$ is an unnormalized likelihood and $(x)_+ = \max(0, x)$. Then, using the idea of data augmentation (Polson & Scott, 2011), we can show the equality

$$\phi'(y|z_d, \boldsymbol{\eta}) = \int_0^\infty \frac{1}{\sqrt{2\pi\omega^y_d}} \exp\left(-\frac{(\omega^y_d + c\tilde{\zeta}^y_d)^2}{2\omega^y_d}\right) d\omega^y_d,$$

where $\tilde{\zeta}^y_d = \kappa^y_d\zeta^y_d - \kappa^y_d f(y, \mathbf{x}_d; z_d, \boldsymbol{\eta})$. Therefore, the conditional distribution $q(\boldsymbol{\eta}^y_k|\mathbf{Z}, \boldsymbol{\eta}^{-y}_k)$ can be expressed as the marginal of the following complete distribution with augmented variables $\boldsymbol{\omega}^y = \{\omega^y_d\}$

$$q(\boldsymbol{\eta}^y_k, \boldsymbol{\omega}^y|\mathbf{Z}, \boldsymbol{\eta}^{-y}_k) \propto p_0(\boldsymbol{\eta}^y_k) \prod_{d:z_d=k} \phi'(y, \omega^y_d|z_d, \boldsymbol{\eta}), \quad (21)$$

where $\phi'(y, \omega^y_d|z_d, \boldsymbol{\eta}) = \frac{1}{\sqrt{2\pi\omega^y_d}} \exp\big(-\frac{(\omega^y_d + c\tilde{\zeta}^y_d)^2}{2\omega^y_d}\big)$. Then we can sample $\boldsymbol{\eta}^y_k$ by iterating over the following two steps and finally dropping $\boldsymbol{\omega}^y$:

1) **For $\boldsymbol{\omega}^y$**: we only need to consider ω^y_d where $z_d = k$. Due to the conditional independence, we can sample each ω^y_d separately from the conditional distribution

$$q(\omega^y_d|\mathbf{Z}, \boldsymbol{\eta}) \propto \mathcal{GIG}\left(\omega^y_d; \frac{1}{2}, 1, (c\tilde{\zeta}^y_d)^2\right),$$

where $\mathcal{GIG}(x; p, a, b) = C(p, a, b)x^{p-1}\exp(-\frac{1}{2}(\frac{b}{x} + ax))$ is a generalized inverse Gaussian distribution (Devroye, 1986) and $C(p, a, b)$ is a normalization constant. Therefore, $(\omega^y_d)^{-1}$ follows an inverse Gaussian distribution

$$q((\omega^y_d)^{-1}|\mathbf{Z}, \boldsymbol{\eta}) = \mathcal{IG}\left((\omega^y_d)^{-1}; \frac{1}{c|\tilde{\zeta}^y_d|}, 1\right), \quad (22)$$

where $\mathcal{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}}\exp(-\frac{b(x-a)^2}{2a^2x})$ for $a, b > 0$. We can efficiently draw samples from an $\mathcal{IG}$ distribution (Michael et al., 1976), with $\mathcal{O}(1)$ time complexity.

2) **For $\boldsymbol{\eta}^y_k$**: this step is to draw the classifier parameter for each active cluster. For the commonly used Gaussian prior, $p_0(\boldsymbol{\eta}^y_k) = \mathcal{N}(0, \nu^2 I)$, we have the conditional distribution

$$q(\boldsymbol{\eta}^y_k|\mathbf{Z}, \boldsymbol{\omega}, \boldsymbol{\gamma}) \propto p_0(\boldsymbol{\eta}^y_k) \prod_{d:z_d=k} \phi'(y, \omega^y_d|z_d, \boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}^y_k, \Lambda^y_k), \quad (23)$$

where mean $\boldsymbol{\lambda}^y_k = \Lambda^y_k(c\sum_d \delta_{z_d,k}\frac{\rho^y_d + c\omega^y_d\kappa^y_d}{\omega^y_d}\mathbf{x}_d)$ and covariance $\Lambda^y_k = (\frac{1}{\nu^2}I + c^2\sum_k \delta_{z_d,k}\frac{\mathbf{x}_d\mathbf{x}_d^\top}{\omega^y_d})^{-1}$.

With the above conditional distributions, we set the initial number of states K_0 to a relatively large value and then randomly initialize $\boldsymbol{\eta}$. Then we construct a Markov chain to iteratively draw samples of $\mathbf{Z}$ using Eq. (17,18) and draw $\boldsymbol{\eta}^y_k$ using the above two-step inner sampler, with an initial condition. In our experiments, we initially set $\boldsymbol{\omega} = 1$ and randomly draw $\mathbf{Z}$ from a K_0 dimensional uniform distribution. In training, we run this Markov chain until convergence (i.e., finished the burn-in stage with M iterations). Then, we draw a sample $\hat{\boldsymbol{\eta}}$ for each component as the final Gibbs classifier to make predictions on testing data.

4.2. Sequential Models using Beam Sampler

Since we can apply a similar method for multiple sequences, here we consider the general problem (14) for one sequence, whose solution is

$$q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\pi}) = \frac{p_0(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\pi})p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\gamma})\phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta})}{\psi(\mathbf{X}, \mathbf{y})},$$

where $\phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta}) = \exp(-2c\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta}))$ is an unnormalized likelihood corresponding to the structured hinge loss (12). Similarly, we can integrate out $\boldsymbol{\gamma}$ by conjugacy and perform collapsed inference on $q(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\pi})$. We can develop an efficient sampler by leveraging the advances in the Beam sampler for iHMMs (Gael et al., 2008)[4]. Specifically, we introduce a set of auxiliary variables $\boldsymbol{\mu}$. Then, we perform the following steps:

For $\boldsymbol{\mu}$: for each time t we draw an auxiliary variable $\mu_t \sim \mathrm{U}(0, \pi_{z_{t-1}, z_t})$.

For Z: since the sequences are conditionally independent given the global variables $(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\eta})$, we can sample the trajectory of each sequence separately. This can be efficiently done with a forward filtering-backward sampling procedure. For the forward filtering, we compute $q(z_t|y_{1:t}, \mu_{1:t})$ using the iterative rule:

$$q(z_t|y_{1:t}, \mu_{1:t}) \propto q(z_t, \mu_t, y_t|y_{1:t-1}, \mu_{1:t-1}) = \phi(y_t|z_t) \sum_{z_{t-1}:\mu_t < \pi_{z_{t-1}, z_t}} q(z_{t-1}|y_{1:t-1}, \mu_{1:t-1}), \quad (24)$$

where $\phi(y_t|z_t) = \exp(-2c \max_y(\Delta_t^y + f(y_t^*, \mathbf{x}_t; z_t, \boldsymbol{\eta}) - f(y, \mathbf{x}_t; z_t, \boldsymbol{\eta})))$ is the unnormalized likelihood for each data point. Then, the backward sampling performs a backward pass where we sample z_t given the sample for z_{t+1}:

$$q(z_t|z_{t+1}, \mathbf{y}, \boldsymbol{\mu}) \propto q(z_t|y_{1:t}, \mu_{1:t}) q(s_{t+1}|s_t, \mu_{t+1}), \quad (25)$$

where $q(s_{t+1}|s_t, \mu_{t+1}) = \pi_{s_t, s_{t+1}} \mathbb{I}(\mu_{t+1} < \pi_{s_t, s_{t+1}})$.

For $\boldsymbol{\eta}$: due to the separability of the discriminant function, the posterior of each classifier weight is $q(\boldsymbol{\eta}_k|\mathbf{Z}, \mathbf{y}) \propto p_0(\boldsymbol{\eta}_k) \prod_{t:z_t=k} \exp(-2c \max_y(\Delta_t^y + \boldsymbol{\eta}_k^\top (g(y_t^*, \mathbf{x}_t) - g(y, \mathbf{x}_t))))$, where $\mathbf{x}_t$ is the t-th segment of the sequence. This step can be done with data augmentation, similar as in the single variable case.

For $\boldsymbol{\pi}, \boldsymbol{\beta}$: these follow from the theory of HDPs. Details can be found in (Teh et al., 2006). We omit for space.

4.3. Prediction

We can apply an iterative algorithm to make predictions for M2iHMM-1. That is, to minimize the latent discriminant function (10) by first sampling the state indicator variable $\mathbf{Z}$ from $q(\mathbf{Z}|\mathbf{X}, \boldsymbol{\pi})$ and then make predictions by (11). Then we can infer $q(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{y}}, \boldsymbol{\pi})$ and make predictions $\hat{\mathbf{y}}$ for the next iteration using the latent discriminant function (10) where we sample $\mathbf{Z}$ from $q(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{y}}, \boldsymbol{\pi})$ instead of $q(\mathbf{Z}|\mathbf{X}, \boldsymbol{\pi})$. In our experiments, we find that doing about 50 iterations is enough for convergence. For M2iHMM-2 we do prediction in a same framework, using $q(\mathbf{Z}|\boldsymbol{\pi})$, $q(\mathbf{Z}|\hat{\mathbf{y}}, \boldsymbol{\pi})$ instead of $q(\mathbf{Z}|\mathbf{X}, \boldsymbol{\pi})$, $q(\mathbf{Z}|\mathbf{X}, \hat{\mathbf{y}}, \boldsymbol{\pi})$ separately. For single variable classification the procedure is similar.

[4]For problem (13), a similar sampler applies.

The Gibbs classifiers only apply a single sample to make predictions, which can be unstable. In our experiments we draw several samples (e.g. 100 samples) and do majority voting to achieve a more stable prediction.

5. Experiments

We now provide empirical studies for both single variable classification and sequential prediction on several synthetic and real data sets. For single variate classification, DP mixtures of Multinomial Logit Model (DPMNL) and iSVM can serve as strong baselines. For dynamic models, since iM2EDM is the most similar model as ours and it has shown superior prediction performance on several data sets (Chatzis, 2013), we use it as a strong baseline. We implemented our models and re-implemented DPMNL and iM2EDM using C++. All the experiments were conducted on an Intel Core i5 3.10GHZ computer with 4.0GB RAM.

Table 1. Classification accuracy (%) and F1 scores (%) on the Parkinsons and Protein data sets.

	PARKINSONS		PROTEIN	
MODEL	ACCURACY	F1 SCORE	ACCURACY	F1 SCORE
MNL	85.6 ± 2.2	79.1 ± 2.8	50.1 ± 0.0	43.5 ± 0.0
LINEAR-SVM	85.3 ± 0.4	78.9 ± 1.5	48.3 ± 0.0	43.2 ± 0.0
RBF-SVM	87.2 ± 2.7	79.9 ± 3.2	53.1 ± 0.0	49.5 ± 0.0
DPMNL	87.7 ± 3.3	82.6 ± 2.5	$\mathbf{56.3} \pm 0.0$	49.5 ± 0.0
ISVM	88.0 ± 1.5	83.5 ± 2.8	54.3 ± 0.0	49.4 ± 0.0
GISVM	$\mathbf{88.9} \pm 1.5$	$\mathbf{85.1} \pm 1.3$	55.8 ± 0.0	$\mathbf{50.1} \pm 0.0$

5.1. Single Variable classification

5.1.1. DATA DESCRIPTION

Parkinsons data: The Parkinsons data set contains 195 instances with 147 positive instances and 48 negative ones. The original data set has 23 features detecting the Parkinsons disease and we extract 10 principal components using PCA, following (Shahbaba & Neal, 2009). We adopt 5-fold cross-validation and report the average performance as well as standard deviations.

Protein data: To recognize structures of Protein, people utilized informative features (i.e., the length of the protein sequence) to predict the folding classes of a protein, which is closely related to its 3D structure. We follow (Ding & Dubchak, 2001) to split the data set into a training set containing 313 instances and a test set consisting of 385 instances. Each instance belongs to one of the 27 folding classes, and is characterized by 21 features.

5.1.2. RESULTS

We compare the average prediction accuracy and F1 scores over linear classifiers such as Multinomial Logit Model (MNL) and SVM using linear kernels (linear-SVM) together with non-linear models such as SVM with RBF kernels (RBF-SVM), DPMNL, iSVM, and GiSVM on the above data sets. For iSVM we set the truncation level $K =$

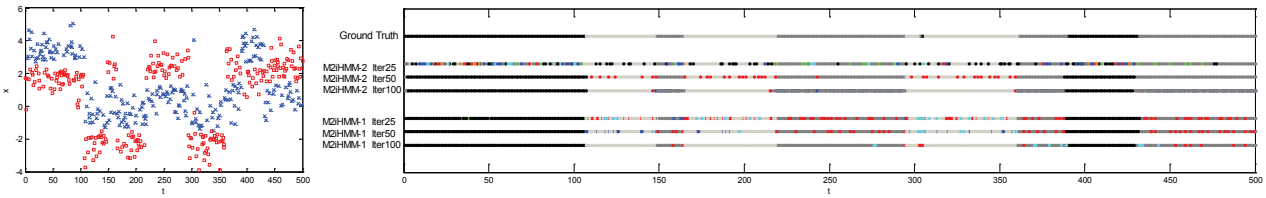


Figure 2. (L) Training data with length 500 in one-dimensional space (markers). The labels (denoted by color/style of markers) are generated by the classifiers w.r.t. the according hidden states. (R) The ground truth states (denoted by different colors on the bar) for the training data and the states recovered by M2iHMM-1 and M2iHMM-2 after 25, 50, 100 iterations separately.

20, and use cross-validation to choose hyper-parameters.

We can see that non-linear models are superior in prediction. Large-margin methods with a mixture-of-experts (i.e., iSVM and GiSVM) can further improve prediction performance in the Parkinsons data set. In both data sets GiSVM shows superior performance than iSVM, and this performance gap may due to the inaccuracy of the truncated mean-field approximation in iSVM. For Parkinsons data we inferred $3 \sim 4$ clusters using GiSVM, which was similar to DPMNL that can detect some structures in data (e.g., heterogeneity of subjects (Shahbaba & Neal, 2009)).

5.2. Sequential Models

5.2.1. Synthetic Data

Training Behavior We generate an observation chain with length 500 from a Gaussian mixture model consisting of three components, whose means are 2.5, 1.5 and -1.5 respectively, while having the same covariance as one. For each observation we set its label by a ground-truth classifier w.r.t. the component it belongs to (i.e., a decision boundary that goes across the mean of the Gaussian component), as shown in Fig. 2(L). The component indicators (or hidden states) are drawn from a Markov chain with a transition matrix $0.01E_3 + 0.98I_3$.[5]. Given the observations and labels, we use M2iHMM-1 and M2iHMM-2 to estimate the classifiers $\boldsymbol{\eta}$ and the hidden states $\mathbf{Z}$. We also estimate the Gaussian components $\boldsymbol{\gamma}$ in M2iHMM-1. In this experiment we set the initial number of states $K_0 = 10$, the HDP concentration hyper-parameters $\alpha_0 = 2, \gamma_0 = 2$, and the large-margin classifier hyper-parameters $c = 1$, $\ell = 1$.[6] The recovered states are shown in Fig. 2(R), where colors in the bars represent the inferred states for the according observations.

We find that samples drawn from M2iHMM-1 and M2iHMM-2 are stable after about 100 iterations (as illustrated in Appendix). The results of the inferred states show some interesting points. First, M2iHMM-2 performs pretty bad in the first 25 iterations while it quickly converges to the ground truth with minor errors (an analysis of convergence is illustrated in the Appendix). Second, by applying a likelihood model, M2iHMM-1 performs fairly good in the first 25 iterations with the information in the observation space. However, because the observations generated by the three states overlap, M2iHMM-1 cannot perfectly recovering all the hidden states with a likelihood model using a moderate weight for the observation likelihood.

[5] E_3 stands for a three-dimensional matrix with all elements equal to one, and I_3 is a three-dimensional identity matrix.

[6] We set a moderate value for c to let the observation likelihood play a relatively strong part in our model.

Table 2. Classification accuracies (%) and time cost for different models in three different synthetic situations.

Model	Set.1(pos.)	Set.2(neg.)	Set.3(vague)	Trn.Time	Tst.Time
SVM	61.9 ± 13.6	64.0 ± 14.5	63.2 ± 14.5	3 ± 0.2s	0.4 ± 0.05s
iM2EDM	67.2 ± 14.2	67.6 ± 13.9	66.7 ± 14.8	164 ± 37s	110 ± 24s
M2iHMM-2	**69.4** ± 12.6	**70.4** ± 12.3	**70.2** ± 12.8	45 ± 7s	19 ± 3s
GiSVM	79.6 ± 16.6	79.1 ± 15.3	82.2 ± 13.9	24 ± 2s	3.6 ± 0.7s
M2iHMM-1	**91.2** ± 5.9	**85.0** ± 9.9	**84.8** ± 10.6	53 ± 10s	15 ± 4s

Testing Behavior Now we show the prediction performance and time cost for different models. Considering the random effect of data, for training we randomly sample 10 data sets in one-dimensional space, and each with an observation chain and corresponding labels with length 500. Similar as above, each observation is drawn from one of three Gaussian components. The mean for each Gaussian component is drawn from $[-3, 3]$, and we set the covariance to one for all the components. Therefore, there is a high probability that some components may largely overlap in the observation space, and it is hard to decide the true number of components barely from the observation space. For each observation we draw its label from a ground-truth classifier w.r.t its component.

For each data set we try three transition matrices when generating the state chains, namely a positive correlated transition matrix $0.05E_3 + 0.85I_3$, a negative correlated transition matrix $0.45E_3 - 0.35I_3$, and a vague transition matrix $0.3E_3 + 0.1I_3$. For testing we generate a latent state chain with length 5,000 for each data set using the same transition matrix. Then we draw the observations and the ground-truth labels. Our task is to predict the labels for each observation. The hyper-parameter settings are the same as the above training behavior experiment. We draw 100 samples from GiSVM, M2iHMM-1 and M2iHMM-2 after 100 iterations and do voting as we stated in Sec. 4.3. For iM2EDM we run 100 variational iterations and use the model in the last iteration for prediction.

The prediction results are shown in Table 2. We compare over linear SVM, GiSVM, iM2EDM, M2iHMM-1,

and M2iHMM-2. Since the data generated is highly nonlinear, the linear classifiers (e.g., SVM) are very ineffctive in prediction. For sequential models, M2iHMM-2 do better both in prediction accuracy and time cost than iM2EDM by capturing the sequential dependencies among data more accurately. By exploring the observation space with a likelihood model, GiSVM could mitigate the over-fitting effect, and partly discover the local linearity in the data. But it still can not capture the sequential dependencies. Over all settings, M2iHMM-1 performs the best, especially on the positive-correlated data. This is rooted from a clearer sequential dependency in the data.

We also compare training time and testing time for different models. Through an efficient sampling scheme, M2iHMM-2 achieves much less time cost than iM2EDM. Also, modeling the likelihood part does not suffer from much additional cost when comparing M2iHMM-1 with M2iHMM-2. Using an efficient beam sampler, M2iHMM-1 samples the latent variable chain by adopting a fast forward-filtering backward-sampling method given the auxiliary variables so the first order Markov dependencies does not bring about a bottleneck in time efficiency comparing with GiSVM, the single variate prediction model.

5.2.2. Human-Activity Recognition

Human-activity recognition in videos with a stationary background was motivated by several applications such as monitoring patients for health care. Recently, the emergence of various devices based on multi-modal sensors has brought about new research topics for combining various source of video streams (e.g., color-depth video streams) to boost the performance of activity recognition (Ni et al., 2011). RGBD-HuDaAct is a home-monitoring human activity recognition data set containing both color and depth video streams (Ni et al., 2011).

In our experiments, we only use the depth video streams. Different from the experiments proposed by Chatzis (2013), which used a subset of data containing five categories of human activities, we use the whole data set containing 12 categories of human activities in 35 video sequences, from which over five million frames summarized in 1,189 samples was extracted. Following (Ni et al., 2011), we sub-sample 702 samples (each contains a sequence of frames for one activity) and discard samples for background activities. For features, we extract the 162-dimension spatio-temporal interest points (STIPS)[7] and generate a code of size $J = 128$, followed by Locality-constrainted linear coding (LLC) (Wang et al., 2011). Finally we do max-pooling w.r.t. a small time step, resulting in 10,624 128-dimension "super-frames" in total, with an average of 15.6 sequential super-frames for each sample. We perform 5-fold cross-validation and report the average performance as well as standard deviations.

[7]Each STIPS feature contains the 3D coordinates (x, y, z), the temporal index t, the scale of the feature point σ, and the HOG, HOF features. (Ni et al., 2011)

Table 3. Classification accuracies (%) and train time for different models in RGBD-HuDaAct data set.

MODEL	ACCURACY	F1 SCORE	TRAIN TIME
SVM	47.9 ± 5.8	45.2 ± 6.0	0.1H
GISVM	48.6 ± 4.5	46.1 ± 7.0	4.5H
IM2EDM	51.2 ± 5.0	48.2 ± 5.2	6.1H
M2IHMM-1	52.0 ± 5.3	48.5 ± 6.1	8.9H
M2IHMM-2	**54.0** ± 5.5	**51.1** ± 6.8	2.3H

We compare over linear SVM, GiSVM, iM2EDM, and M2iHMMs. For models based on Gibbs classifiers we set K_0=20 and run 300 iterations. Then we do prediction with the final Gibbs classifier, as mentioned in Sec. 4.3. While for iM2EDM we set the truncation level K=20 and run 100 iterations for training. All the models converge in our experiments. The prediction and time cost results was demonstrated in Table 3. The sequential models perform significantly better than the single variable models, due to their ability to capture the sequential dependencies among data. Applying a likelihood model does not help in prediction because the observation space does not provide informative sequential correlations. M2iHMM-2 performs the best both in prediction accuracy and in time efficiency. Finally, M2iHMMs inferred $6 \sim 8$ clusters, which can reflect the complex structures in the data.

6. Conclusions and Future Work

We present max-margin infinite HMMs that explore max-margin discriminative learning on iHMMs for sequential prediction. By introducing Gibbs classifiers and data augmentation representations, we develop truncation-free and assumption-free MCMC algorithms with efficient beam samplers. Empirical results have justified superior performance of our models on both sequential prediction and single variate classification tasks than other competitors.

For future work, we can explore more complicated structures among data (e.g., tree structures) by models that scale potentially to an infinite extent (Blei et al., 2010). Another exciting direction is to develop small-variance asymptotics for our models that admit very fast k-means like estimation algorithms (Campbell et al., 2013).

Acknowledgments

The work is supported by the National Basic Research Program of China (Nos. 2013CB329403, 2012CB316301), National Natural Science Foundation of China (Nos. 61322308, 61332007), Tsinghua University Initiative Scientific Research Program No. 20121088071, and a Microsoft Research Asia Research Fund No. 20123000007.

References

Altun, Y., Tsochantaridis, I., and Hoffman, T. Hidden Markov support vector machines. In *ICML*, 2004.

Antoniak, C.E. Mixture of Dirichlet process with applications to Bayesian nonparametric problems. *Annals of Stats*, (273):1152–1174, 1974.

Beal, M.J., Ghahramani, Z., and Rasmussen, C.E. The infinite hidden markov model. In *NIPS*, 2001.

Blei, D.M., Griffith, T., and Jordan, M.I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57 (2), 2010.

Campbell, T., Liu, M., Kulis, B., and How, J. Dynamic clustering via asymptotics of the Dependent dirichlet process mixture. In *NIPS*, 2013.

Chatzis, S. Infinite Markov switching maximum entropy discriminantion machines. In *ICML*, 2013.

Collobert, R., Bengio, S., and Bengio, Y. A parallel mixture of SVMs for very large scale problems. In *NIPS*, 2002.

Crammer, K. and Singer, Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2: 265–292, 2001.

Devroye, L. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

Ding, C. and Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.

Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Annals of Stats*, (1):209–230, 1973.

Fu, Z., Robles-Kelly, A., and Zhou, J. Mixing linear SVMs for nonlinear classification. *IEEE Trans. on Neural Networks*, 21(12):1963–1975, 2010.

Gael, J., Saatci, Y., Teh, Y.W., and Ghahramani, Z. Beam sampling for the infinite hidden Markov model. In *ICML*, 2008.

Germain, P., Lacasse, A., and Laviolette, F. PAC-Bayesian learning of linear classifiers. In *ICML*, 2009.

Hannah, L.A., Blei, D.M., and Powell, W.B. Dirichlet process mixtures of generalized linear models. Technical report, 2010.

Jaakkola, T., Meila, M., and Jebara, T. Maximum entropy discrimination. In *NIPS*, 1999.

Langford, J. and Shawe-Taylor, J. PAC-Bayes & margins. In *NIPS*, 2003.

Michael, John R., Schucany, William R., and Haas, Roy W. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

Ni, B., Wang, G., and Moulin, P. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *ICCV Workshops*, 2011.

Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145C–158, 1995.

Pitman, J. Combinatorial stochastic processes. Technical report, Department of Statistics, University of California at Berkeley, 2002.

Polson, N.G. and Scott, S.L. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–24, 2011.

Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.

Scott, S.L. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97:337–351, 2002.

Sha, F. and Saul, L.K. Large margin hidden Markov models for automatic speech recognition. In *NIPS*, 2006.

Shahbaba, B. and Neal, R. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

Tanner, M.A. and Wong, W.-H. The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

Taskar, B., Guestrin, C., and Koller, D. Max-margin Markov networks. In *NIPS*, 2003.

Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. Hierarchical Dirichlet process. In *NIPS*, 2006.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. Locality-constrainted linear coding for image classification. In *CVPR*, 2011.

Zhu, J., Chen, N., and Xing, E.P. Infinite SVM: a Dirichlet process mixture of large-margin kernel machines. In *ICML*, 2011.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with fast sampling algorithms. In *ICML*, 2013.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with data augmentation. *JMLR (to appear)*, 2014.