Spectral Learning for Mixture of Markov Models

Y. Cem Sübakan
Dept. of Computer Science
Univ. of Illinois at Urbana-Champaign
Urbana, IL 61801
subakan2@illinois.edu

Barış Kurt[1], **A. Taylan Cemgil**[1], **Bülent Sankur**[2]
Computer Engineering[1], Electrical Engineering[2],
Boğaziçi University
İstanbul, Bebek 34342
{baris.kurt, taylan.cemgil, bulent.sankur}@boun.edu.tr

Abstract

In this paper, we develop two spectral methods for mixture of Markov models. First, we show that to derive spectral learning algorithm for a mixture of Markov models, we need to use moments up to order five. To reduce the sample complexity, we propose an alternative scheme, which is based on learning a mixture of Dirichlet distributions. We experimentally show that the latter approach outperforms the regular spectral learning algorithm for mixture of Markov models, and expectation maximization algorithms for mixture of Markov and Dirichlet models.

1 Introduction

Spectral learning methods have recently become popular in machine learning community due to their ability to learn latent variable models in computationally efficient and local optima-free fashion. The prominent examples include spectral learning for such cases as, inference in Hidden Markov Models [1], inference in arbitrary latent trees and latent junction trees [2, 3], parameter estimation in mixture models and Hidden Markov Models [4, 5].

While deriving a spectral learning algorithm for mixture of Markov models following this methodology, we found that the direct application of the approach in [4] necessitates to use observable moments of up to order five. To overcome this drawback, we propose an alternative scheme where we consider the posterior distribution of the model parameters: In the mixture of Markov models case, if we assume a uniform Dirichlet prior, the posterior distribution of the transition matrices becomes a Dirichlet distribution with empirical transition counts as parameters. So, we can treat a normalized sufficient statistics matrix of a sequence as a sample from this Dirichlet posterior and, learn a mixture of Dirichlet distributions for clustering sequences.

Experiments on synthetic data show that spectral learning of mixture of Dirichlet distributions outperforms the conventional spectral learning approach and expectation maximization algorithms on mixture of Markov and mixture of Dirichlet models.

2 Learning Mixture of Markov Models

A mixture of Markov models is defined by $\{A_{1:K}, \pi\}$ where A_k is the $L \times L$ transition probabilities matrix for the k^{th} Markov model and π is the vector of mixing proportions. The initial state distributions are omitted for the sake of simplicity.

The likelihood of observing a sequence $x_n = (x_{1,n}, x_{2,n}, \ldots, x_{T_n,n})$ of length T_n is defined as:

$$p(x_n|A_{1:K}) = \sum_{k=1}^{K} p(h_n = k) \prod_{t=1}^{T_n} p(x_{t,n}|x_{t-1,n}, h_n = k) = \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T_n} \prod_{l_1=1}^{L} \prod_{l_2=1}^{L} A_{k,l_1,l_2}^{[x_{t,n}=l_1][x_{t-1,n}=l_2]} \tag{1}$$

where, $h_n \in \{1, 2, \ldots, K\}$ is the latent cluster indicator of the sequence x_n. Given N observation sequences $\mathbf{x}_{1:N} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, to estimate the model parameters $A_{1:K}$, one can maximize the likelihood:

$$A^*_{1:K} = \text{argmax}_{A_{1:K}} p(\mathbf{x}_{1:N}|A_{1:K}) = \text{argmax}_{A_{1:K}} \sum_{h_{1:N}} \prod_{n=1}^{N} p(\mathbf{x}_n|h_n, A_{1:K}) \tag{2}$$

which is intractable since the likelihood is defined with a summation over all possible combinations of $h_{1:N}$. The conventional way is to iteratively maximize an EM lower bound.

$$Q(A^{old}_{1:K}, A^{new}_{1:K}) = \mathbb{E}_{p(h_{1:N}|x_{1:N}, A^{old}_{1:K})}[\log(p(\mathbf{x}_{1:N}, h_{1:N}|A^{new}_{1:K}))] \tag{3}$$

However, this approach requires clever initialization for $A_{1:K}$ as the optimization problem does not have unique solution. Alternatively a method of moments based learning algorithm for mixture of Markov models can be derived following the methodology in [4]. The trick is to express the moments of the distribution as a matrix multiplication (or possibly tensor as in our case), so that an eigen-decomposition can be applied. The latter reveals information about the model parameters, and they can be obtained using a function of some observable moments.

Directly applying this approach would require the usage of a fifth order observable moments. (Due to space constraint we do not include the derivation[1] in the paper.) Although the algorithm is theoretically sane, obviously in practice accurate estimation of fifth order moments requires excessive data. In the next section, we propose an alternative scheme, based on learning mixture of Dirichlet distributions which enables us to reduce the sample complexity.

3 Learning Mixture of Dirichlet Distributions

Suppose we place a prior distribution for transition matrices on the class conditional likelihood $p(\mathbf{x}_n|A_{h_n}, h_n)$. Placing a Dirichlet prior $p(A_{h_n}) \sim \text{Dirichlet}(\beta, \ldots, \beta)$ would result in a Dirichlet posterior:

$$\begin{aligned} p(A_{h_n}|\mathbf{x}_n, h_n) &\propto \quad p(\mathbf{x}_n|A_{h_n}, h_n)p(A_{h_n}) \\ &\propto \left(\prod_{t=1}^{T_n}\prod_{l_1=1}^{L}\prod_{l_2=1}^{L} A_{h_n,l_1,l_2}^{[x_{t,n}=l_1][x_{t-1,n}=l_2]}\right)\prod_{l_1=1}^{L}\prod_{l_2=1}^{L} A_{h_n,l_1,l_2}^{\beta-1} \propto \prod_{l_1=1}^{L}\prod_{l_2=1}^{L} A_{h_n,l_1,l_2}^{c^n_{l_1,l_2}+\beta-1} \\ &= \quad \text{Dirichlet}(c^n_{1,1}+\beta-1, c^n_{1,2}+\beta-1, \ldots, c^n_{L,L}+\beta-1) \end{aligned} \tag{4}$$

where, $c^n_{l_1,l_2}$ stores the state transition counts of sequence $\mathbf{x}_n$. So, we can indeed characterize a sequence generated by a Markov model with a Dirichlet distribution, since Dirichlet distribution is the posterior of the transition matrix A_{h_n}. Setting $\beta = 1$ (having a uniform prior), we see that the posterior distribution becomes; $p(A_{h_n}|\mathbf{x}_n, h_n) = \text{Dirichlet}(c^n_{1,1}, c^n_{1,2}, \ldots, c^n_{L,L})$. Therefore, we can treat a normalized state transition count matrix as a sample from the posterior of the transition matrix. Hence, we can effectively cluster sequences by using normalized second order statistics matrices instead of the sequences themselves. This would be equivalent to learning a mixture of Dirichlet distributions, instead of learning a mixture of Markov models.

A spectral learning algorithm for a mixture of Dirichlet distributions would be simpler compared to directly learning a mixture of Markov models, since the former requires to estimate a second order moment (normalized transition count matrix) whereas the latter requires a fifth order moment from the sequences. Let us denote an observed normalized state transition count (vectorized) matrix by $s_n \in \mathbb{R}^{(L^2)\times 1}$. Note that $s_{n,l_1,l_2} = c^n_{l_1,l_2}/(\sum_{l_1,l_2} c^n_{l_1,l_2})$. In mixture of Dirichlet distributions, the likelihood of an observation s_n is defined as follows:

$$p(s_n|\alpha) = \sum_{k=1}^{K} p(h_n = k)\text{Dirichlet}(s_n; \alpha_k) \tag{5}$$

where, $\alpha \in \mathbb{R}^{(L^2)\times K}$ is the matrix that stores the Dirichlet parameters in its columns, and α_k is the k'th column of α. The empirical moments and the parameters are connected through the following

[1] The derivation can be found from http://mazeofamazement.files.wordpress.com/2010/08/mmarkovsupplement1.pdf

equations[2]:

$$m := \mathbb{E}[s]; \quad t^l \in \mathbb{R}^{L^2}, \quad t_i^l := \mathbb{E}[s_i^2] - \frac{1}{\alpha_0+1}\mathbb{E}[s_i] \quad if \quad i = l, \quad t_i^l := \mathbb{E}[s_l s_i] \quad if \quad i \neq l$$

$$M_2 := \mathbb{E}[s \otimes s] - \frac{1}{\alpha_0+1}\text{diag}(m)$$

$$M_3 := \mathbb{E}[s \otimes s \otimes s] - \frac{1}{\alpha_0+2}\left(\sum_{l=1}^{L^2}(e_l \otimes e_l \otimes t^l) + (e_l \otimes t^l \otimes e_l) + (t^l \otimes e_l \otimes e_l)\right) - \frac{2}{(\alpha_0+1)(\alpha_0+2)}\left(\sum_{l=1}^{L^2} m_l(e_l \otimes e_l \otimes e_l)\right)$$

then,

$$M_2 = \frac{1}{\alpha_0(\alpha_0+1)}\sum_{k=1}^{K}\pi_k\left(\alpha_k \otimes \alpha_k\right), \quad M_3 = \frac{1}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\sum_{k=1}^{K}\pi_k\left(\alpha_k \otimes \alpha_k \otimes \alpha_k\right)$$

where, $\alpha_0 = \sum_{k=1}^{K}\alpha_k$, $e_{1:L^2}$ is the canonical basis for $\mathbb{R}^{L^2}$, and $\otimes$ is the outer product. Then, the model parameters can be estimated (up to a scaling factor) using eigenvalue decomposition as in [4] or by orthogonal tensor decomposition in [5]. Note that we have assumed that precision parameter is known a-priori and shared in all clusters. The overall procedure is given in Algorithm 1.

Algorithm 1 Sequence clustering via spectral learning of mixture of Dirichlet distributions.

Input: Sequences $\mathbf{x}_{1:N}$
Output: Clustering assignments $\widehat{h}_{1:N}$
1. Extract normalized transition counts $s_{1:N}$ from $\mathbf{x}_{1:N}$.
2. Compute M_2 and M_3.
3. Estimate $\widehat{\alpha}$ (up to a scaling factor) via spectral learning.
4. Output the cluster assignments $\widehat{h}_n = \text{argmax}_k \text{Dirichlet}(s_n, \widehat{\alpha}_k), \forall n \in \{1, \ldots N\}$.

4 Experimental Results

We generated 100 data sets where each set is composed of 60 sequences which are generated from a mixture of 3 Markov models. For each data set, the prior cluster probabilities $p(h_n)$ and transition matrices $A_{1:3}$ are generated randomly.

We compare the clustering accuracies of the spectral learning algorithms in sections 2 and 3 and the EM algorithms for mixture of Markov models and mixture of Dirichlet models. We define the clustering accuracy as the ratio of correct clustering assignments, which is calculated via resolving the permutation ambiguity of the clustering result: The estimated clustering assignments are compared with the true assignments for all possible permutations of cluster identifiers ($K!$ permutations for K clusters), and the maximum ratio is chosen.

In the experimental setting, each sequence has a length of 100000 time instances. In order to see the effect of the number of data instances on clustering accuracy, we redo the experiment for varying sequence lengths. The results are shown in Figure 1. (We redo the experiment by having a sequence dataset comprising of sequences of length shown on the x-axis.)

We see that mixture of Dirichlet distributions yield the highest clustering accuracies for all sequence lengths. Also, we observe that when we have large number of samples, spectral algorithms tend to give higher clustering accuracies compared to their EM counterparts. In the next set of experiments, we investigate the effect of changing L (cardinality of observations) and cluster similarity on clustering accuracy. In cluster similarity experiment, for $K = 3$, the transition matrices are generated as $A_k = (1-\lambda)\tilde{A}_0 + \lambda\tilde{A}_k$, where $\tilde{A}_0, \tilde{A}_1, \tilde{A}_2, \tilde{A}_3 \sim$ Dirichlet(1,...,1). We repeat both experiments for multiple datasets, and report the average clustering accuracy. The results for both experiments, for differing sequence lengths are given in Figure 2.

[2]The derivation can be found from http://mazeofamazement.files.wordpress.com/2010/08/mmarkovsupplement1.pdf

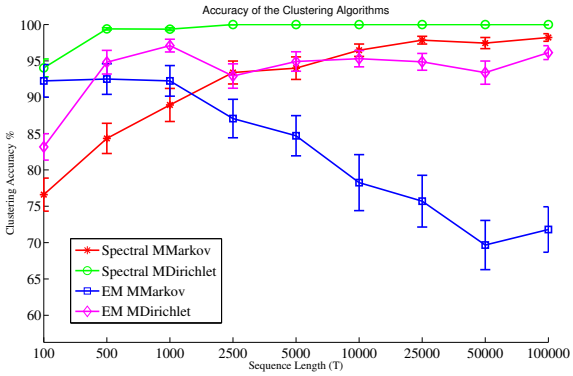
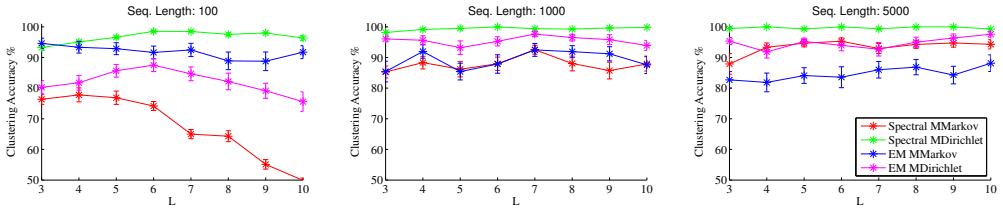
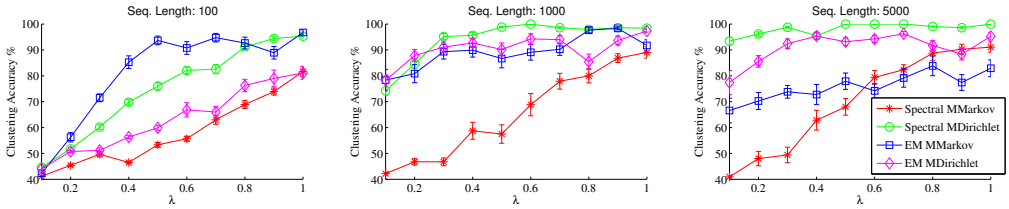


Figure 1: Comparison of clustering accuracies on synthetic data for differing sequence lengths



(a) Effect of changing L experiment



(b) Effect of cluster similarity experiment

Figure 2: Comparison of clustering accuracies on synthetic data for differing sequence lengths

We observe that number of states L has the least significant effect on spectral mixture of Dirichlet algorithm, amongst all algorithms. In experiments with short sequences, we see that the spectral learning for mixture of Markov models is the most sensitive algorithm to increasing L. As expected, we also see that all algorithms become less sensitive to L as sequence length increases. In cluster similarity experiment, we observe that if there is enough data available, mixture of Dirichlet algorithm yields high accuracy, even when $\lambda = 0.1$.

5 Conclusion and Future Work

We have proposed a simple, fast and high performance spectral learning algorithm for clustering sequences. Our approach helps to reduce the sample complexity of the spectral learning algorithm as we have experimentally showed that our approach outperforms regular spectral learning of mixture of Markov models on all sample rates. Furthermore, our algorithm outperforms expectation maximization algorithms for mixture of Markov models and mixture of Dirichlet distributions. As future work, we may investigate a similar procedure which is based on learning the posterior distribution of parameters for deriving method of moments based learning algorithms for more sophisticated latent variable models with temporally connected observations.

References

[1] Hsu, D., S. M. Kakade and T. Zhang, "A Spectral Algorithm for Learning Hidden Markov Models A Spectral Algorithm for Learning Hidden Markov Models", *Journal of Computer and System Sciences*, , No. 1460-1480, 2009.

[2] Parikh, A., L. Song and E. Xing, "A spectral algorithm for latent tree graphical models", *ICML*, 2011.

[3] Parikh, A. P., L. Song, M. Ishteva, G. Teodoru and E. P. Xing, "A Spectral Algorithm for Latent Junction Trees", *The 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[4] Anandkumar, A., D. Hsu and S. Kakade, "A Method of Moments for Mixture Models and Hidden Markov Models", *COLT*, 2012.

[5] Anandkumar, A., R. Ge, D. Hsu, S. Kakade and M. Telgarsky, "Tensor Decompositions for Learning Latent Variable Models", *arXiv:1210.7559v2*, 2012.