Spectral Learning for Mixture of Markov Models

Y. Cem Sübakan¹, Barış Kurt², A. Taylan Cemgil², Bülent Sankur³

¹ University of Illinois at Urbana-Champaign, Computer Science
² Boğaziçi University, Computer Engineering
³ Boğaziçi University, Electrical & Electronics Engineering

Problem statement: Sequence clustering by spectral learning for Mixture of Markov models.

Contribution: Regular spectral learning algorithms for latent variable models require fifth order moment. We reduce the sample complexity by learning mixtures of Dirichlet distributions.

Conclusion: We experimentally show the superiority of our approach compared to EM and regular spectral learning.

The graphical model is as follows:

![Graphical Model](image)

- Observation at sequence \( x_n \), time \( t: x_{tn} \).
- Cluster indicator of sequence \( \alpha_n \).
- Transition distribution of cluster \( \beta_k \).

The likelihood of observing a sequence \( x_n = (x_{tn})_{1 \leq t \leq L} \) of length \( L \) is defined as:

\[
 p(x_n|A_k) = \prod_{i=1}^{L} p(x_{i+1}|x_i, A_k, h_k)
\]

\[
 = \prod_{i=1}^{L} \sum_{h_k} A_{k}(x_{i+1}|x_i, h_k)
\]

The ultimate learning goal is to estimate the cluster assignments \( h_k \) given sequences \( x_n \).

2. Spectral Learning of Mixture of Markov Models

Learning Strategy: Learn transition matrices \( \{A_k\}_1 \) of latent variable models, then learn the assignments \( h_k \).

Claim: The model parameters \( \{A_k\}_1 \) can be uniquely identified using fifth and fourth order moments:

\[
 B_{jk} = \sum_{i=1}^{L} A_{j}(x_{i+1}|x_i, h_k) \quad B_{jk} = \sum_{i=1}^{L} A_{j}(x_{i+1}|x_i, h_k)
\]

Drawback: This spectral learning approach requires moments up to order five.

3. Spectral Learning of Mixture of Dirichlet Distributions

Learning Strategy: Alternatively, one can learn a mixture of posterior distributions of transition matrices, which is mixture of Dirichlet distributions, to estimate \( h_k \).

Using conjugate Dirichlet prior, the posterior is also Dirichlet:

\[
p(\alpha_k|x_n, A_k) \propto p(\alpha_k)p(x_n|\alpha_k, A_k) = \prod_{i=1}^{L} \sum_{h_k} \alpha_i(k|1)-1 \prod_{l \neq k} \alpha_l(k|1)-1
\]

\[
 = \prod_{i=1}^{L} \sum_{h_k} \alpha_i(k|1)-1 \prod_{l \neq k} \alpha_l(k|1)-1
\]

\[
 = \text{Dirichlet}(\alpha_1 + \beta - 1, \ldots, \alpha_K + \beta - 1)
\]

where, \( \alpha_i(k) \) stores the state transition counts of \( x_n \).

Setting \( \beta = 1 \) (having a uniform prior), the posterior distribution becomes:

\[
p(\alpha_k|x_n, A_k) \propto p(\alpha_k)p(x_n|\alpha_k, A_k) = \prod_{i=1}^{L} \sum_{h_k} \alpha_i(k|1)-1 \prod_{l \neq k} \alpha_l(k|1)-1
\]

We treat the normalized state transition count matrices \( \alpha_{i,l}^{k} = \alpha_i(k)/\sum_{l} \alpha_l(k) \) as a sample from the posterior of the transition matrix.

So, the graphical model becomes:

![Graphical Model](image)

- Observations, normalized transition counts: \( x_n \).
- Dirichlet parameters of cluster: \( \alpha_n \).

Claim: The posterior parameters of each cluster and observable moments are related as the following:

\[
m = E[\alpha_k], \quad \mu = E[\alpha_k - 1]
\]

\[
 \mu = E[\alpha_k] h_k - E[\alpha_k|\alpha_k - 1]
\]

\[
 \mu = E[\alpha_k|\alpha_k - 1] - (\alpha_k - 1) \text{diag}(\alpha_k)
\]

Then,

\[
 M_d = \sum_{k_1} \mu(k_1, k_2) \cdot \mu(k_2, k_3) + \mu(k_1, k_3)
\]

where, \( \mu = \sum_{l} \alpha_l(k) \) is the canonical basis for \( \mathbb{R}^{|K|} \) and \( \otimes \) is the outer product operator.

Having the parameters in the symmetric tensor form, we can apply the existing spectral learning procedures to estimate \( \alpha_n \).

4. Experimental Results

- We generated 100 data sets. Each set is composed of 60 sequences, with \( K = 3 \).
- The prior cluster probabilities \( p(h_k) \) and transition matrices \( \{A_k\} \) are generated randomly.

First Experiment:

- Comparison of clustering accuracies of the spectral learning and EM algorithms (4 algorithms).
- Results for varying sequence lengths are shown in Figure 3.

Second Experiment:

- We next investigate the effect of changing \( L \) (cardinality of observations). Results for differing \( L \) are given in Figure 4.

- Number of states \( L \), has the least significant effect on spectral mixture of Dirichlet algorithms.
- In experiments with short sequences, the spectral learning for mixture of Markov models is the most sensitive algorithm to increasing \( L \).
- All algorithms become less sensitive to \( L \) as sequence length increases.

Third Experiment:

- Next, we investigate cluster similarity on clustering accuracy.
- For \( K = 3 \), transition matrices are generated as \( \tilde{A}_k = (1 - \lambda)A_1 + \lambda A_2 \), where \( A_1, A_2, \ldots \), Dirichlet(1,...,1). Results for differing \( \lambda \) are given in Figure 5.

- If there is enough data available, mixture of Dirichlet algorithms yields high accuracy, even when \( \lambda = 0.1 \).

5. Conclusions

- Conclusion: Experimental results suggest that proposed method outperforms EM and regular spectral learning in several regimes.
- Future Work: Application of the algorithm on real-world applications.