Factorial Hidden Markov Model

ENNAFII Oussama ENS Cachan

Introduction

- Hidden Markov Models (HMM) are widely used as learning models for time series data (such an application is speech recognition modeling).
- A generalisation of this model is what we call factorial(distributed HMM).
- In this framework, the hidden state variable is actually a vector of multiple state variables.
- A practical example where it comes handy to use this model, is when the sound recorded comes from multiple sources and we want to recognize separatly the two sounds.

The Model

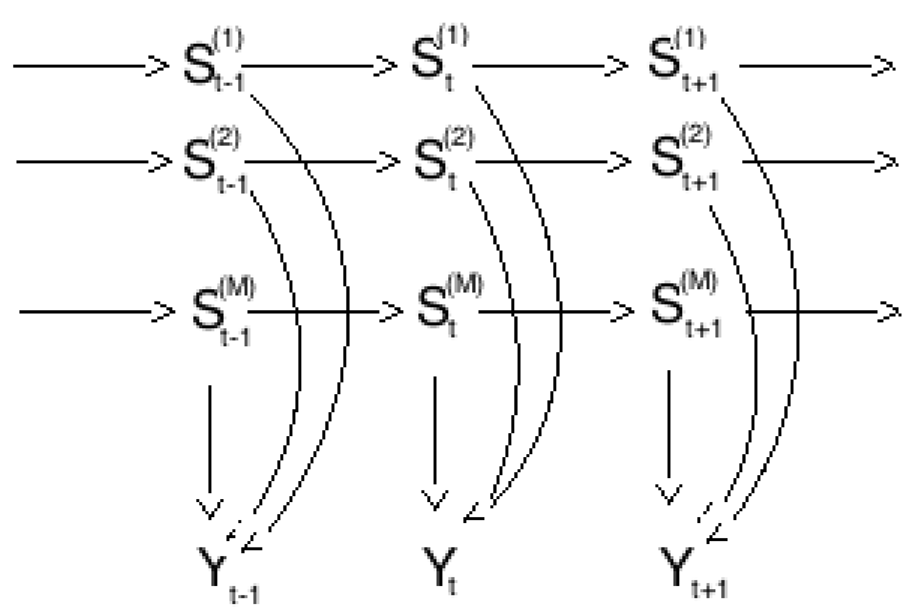

Figure 1: The FHMM represented as a DAG

The factorial model is represented by:

$$P((S_t, Y_t), \forall t) = P(S_1)P(Y_1|S_1)$$

$$\prod_{t=2,\dots,T} P(S_t|S_{t-1})P(Y_t|S_t)$$

Assumptions

The DAG associated to FHMM means that:

$$P(S_t|S_{t-1}) = \prod_{m=1,\dots,M} P(S_t^{(m)}|S_{t-1}^{(m)}) \qquad (1)$$

We further suppose that the emission probabilities are guassian with linear combinations of the means of each substate:

$$P(Y_t|S_t) = N(Y_t, \sum_{m=1\dots M} W^{(m)} S_t^{(m)}, C) \qquad (2)$$

Inference

We use the EM algorithm to infere the hidden variables and to learn the parameters. In the M step we get:

$$W^{new} = (\sum_{t=1,\dots,T} Y_t < S_t^* >)(\sum_{t=1,\dots,T} < S_t S_t^* >)^\dagger \qquad (3)$$

$$\pi^{(m)\ new} =< S_t^{(m)} > \qquad (4)$$

$$P_{i,j}^{(m)\ new} = \frac{\sum_{t=2,\dots,T} < S_{t,i}^{(m)} S_{t,j}^{(m)} >}{\sum_{t=2,\dots,T} < S_{t,j}^{(m)} >} \qquad (5)$$

$$C^{new} = \frac{1}{T}(\sum_{t=1..T} Y_t Y_t^* - \sum_{t,m} W^{(m)} < S_t^{(m)} > Y_t^*) \qquad (6)$$

The Exact Inference:

We get through Forward-Backward algorithm:

$$P(S_t|\{Y_\tau\}_1^T, \phi) = \frac{\alpha_t(S_t)\beta_t(S_t)}{\sum_{S_t} \alpha_t(S_t)\beta_t(S_t)}$$

We deduce the means from this probability distribution.This methods complexity is: $O(TMK^{M+1})$

Variational Inference:

In the variational inference, instead of calculating $P(S_t|\{Y_\tau\}_1^T, \phi)$ (which is intractable), we approach it with parametrerized probabilities. There are two ways to do it:
The completely factorised distribution, where all the states are decoupled multinomial distributions:

$$Q((S_t)_t) = \prod_{t=1,\dots,T} \prod_{m=1,\dots,M} \prod_{S_t=1,\dots,K} (\theta_{t,k}^{(m)})^{S_t^{(m)}}$$

The minimization of the Kullback-Leibler divergence in this case yields:

$$\theta_t^{(m)} = f(\phi, \theta_{t-1}^{(m)}, \theta_{t+1}^{(m)}, (\theta_t^{(n)})_{n\neq m})$$

This scheme converges rapidly (≈ 10 iterations). The means are easy to get formaly since each state variable is a multinomial distribution .

Structured variational Infe

The structured distribution:

$$Q(S_t) \propto \prod_{m=1..M} Q(S_1^{(m)}|h) \prod_{t=1..T} Q(S_t^{(m)}|S_{t-1}^{(m)}, h)$$

with:

$$Q(S_t^{(m)}|S_{t-1}^{(m)}, h) = \prod_{k=1..K} (h_{t,k}^{(m)}) \prod_{j=1..K} (P_{k,j}^{(m)})^{S_{t-1}^{(m)}}$$

Now h has the role of emission probabilities. We keep the Markov property while decoupling the chains. We get another fixed point equation:

$$h_t^{(m)} = g(\phi, (h_t^{(n)})_{n\neq m})$$

Once the scheme converges, we use the forward backward algorithm. The computational complexity of the last step is : $O(TMK^2)$

Numerical results :

We generate artificially unidimensional data using a Factorial HMM with random transition matrices, priors and means. We set $C = 0.005$,$K = 2$ and $M = 3$. We repeat the process 20 times, we get this table reprensenting the loglikelihood per time (divided by T):

Algorithm	Training Set	Test set
Naive	1.28	1.11
Exact	2.32	1.43
Variational	1.56	1.29

Conclusion :

- The direct exact E step is intractable when there are too many sources (for big M)
- The Variational E step is less greedy w.r.t computational complexity. It is a nice tractable method comparing to the naive or the exact ones.