가우시안 혼합 출력 HMM을 위한 변분 베이지안 방법

오장민[O] 장병탁
서울대학교 컴퓨터 공학부
{jmoh[O], btzhang}@bi.snu.ac.kr

Variational Bayesian Methods for Learning HMM with Mixture of Gaussian Outputs

Jangmin O[0] and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

은닉 마코프 모델은 이산 동역학을 표현할 수 있는 확률 모형이다. 우도 함수 최적화를 수행하는 전통적인 Baum-Welch 학습 알고리즘은 국소해로 수렴하기 쉬우며, 우도함수의 특성상 복잡한 모델을 선호하는 바이어스가 존재한다. 베이지안 프레임워크에서는 파라미터를 랜덤 변수로 보고 이에 대한 사후 확률 분포를 추정하여 이 문제를 해결할 수 있다. 본 논문에서는 베이지안 추정을 위한 결정론적 근사화 기법인 변분 베이지안 방법을 이용, 출력 노드에 가우시안 혼합 노드를 지니는 일반화된 HMM의 추론 방법을 유도한다. 인공 데이터에 대한 실험을 통해, 본 방법이 효과적인 HMM 학습을 수행할 수 있음을 보인다.

1. 서 론

은닉 마코프 모델 (HMM: Hidden Markov Model)은 이산 시간 간격의 동역학을 모델링 하는 모델로서, 자연언어처리, 음성인식, 필기체 인식, 금융공학 등에서 널리 활용되고 있다 [5]. HMM은 $\theta = (A, \pi, \Omega)$의 파라미터를 지닌다. A는 상태간 전이 확률이며, π는 초기 상태의 확률이다. Ω는 출력 노드의 확률 분포 모델 파라미터이다. HMM 하에서, 관측 데이터 $y_{1:T}$와 은닉 상태 시퀀스 $s_{1:T}$에 대한우도 함수는 다음과 같다.

$$p(s_{1:T}, y_{1:T}) = p(s_1)p(y_1 \mid s_1)\prod_{t=2}^{T} p(s_t \mid s_{t-1})p(y_t \mid s_t)$$

HMM에서의 추론은 관측 데이터의 evidence 인 $p(y_{1:T}|\theta)$와 은닉 상태의 사후 확률 $p(s_{1:T} \mid y_{1:T}, \theta)$을 추정하는 것이다. 주어진 파라미터 하에서의 추론으로는 forward-backward 기법이 전형적인 방법이다. HMM의 전통적인 학습은, 학습데이터 L개에 대한 로그 우도 함수

$$\log \sum_{l=1}^{L} p(y_{1:T}^{(l)} \mid s_{1:T}^{(l)}, \theta)$$

를 최대화 하는 것으로 EM 알고리즘인 Baum-Welch 기법을 사용한다 [5].

최대 우도 기법은 통계학의 점근적 최적화 특성에도 불구하고 지역해 수렴 및, 복잡한 모델에 대한 편향 등의 문제가 존재한다. 베이지안 추론은 파라미터에 대한 사후 확률 분포를 추정하여 $p(y_{1:T}) = \int p(y_{1:T} \mid \theta) dF_\theta$과 같은 적분 계산을 가능하게 한다. HMM의 경우, 베이지안 추론시의 분석적인 적분이 어렵기 때문에, 결정론적인 근사화 기법인 변분 베이지안 방법이 적용된 사례가 있다 [1, 3, 4, 6]. [1]은 출력 노드가 이산 노드일 때의 변분 베이지안 방법을 유도하였고, [6]는 가우시안 노드인 경우를 유도하였다. 우리는 가우시안 혼합 노드를 출력 노드로 지니는 일반화된 HMM의 베이지안 추론을 유도한다.

논문의 구성은 다음과 같다. 2장에서는 베이지안 HMM을 소개하고, 근사화 알고리즘의 필요성을 기술한다. 3장에서는 변분 방법의 소개 및, 그에 의한 HMM 추론 및 학습 과정을 유도한다. 4장에서 간단한 인공 데이터에 대한 실험 결과를 보이고, 5장에서 결론을 맺는다.

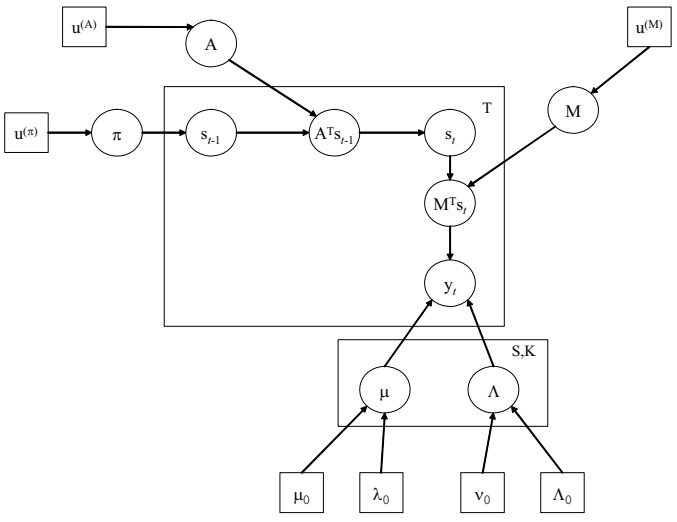

그림 1. 베이지안 HMM의 그래프 표현

2. 베이지안 HMM

그림 1은 베이지안 HMM의 그래프 표현이다. 원 내의 변수는 분포, 사각형 내의 변수는 파라미터이다. 플레이트는 반복을 의미한다. 상태의 수를 S, 출력 혼합 M의 수를 K라고 하자. 출력은 P차원의 가우시안 혼합 노드로서, S×K개 만큼 존재한다. 이 때, A, π, M은 디리클렛 분포이다. μ는 다변수 가우시안 분포, Λ는 위샤트 분포이다. 만일, $p(y_{1:T})$를 계산하고자 한다면, 베이지안 관점에서는,

$$p(y_{1:T}) = \int p(\theta, s_{1:T}) p(s_{1:T}, y_{1:T} \mid \theta) d\theta ds_{1:T}$$

과 같다. 보통의 경우 적분을 수행하는 계산 복잡도가 상당히 크다. 이 경우 베이지안 추론에서는 주로 MCMC 같은 샘플링 기법에 의존한다. 비록 MCMC 기법이 이론적, 실제적으로 강력한 기법이지만, 수행 시간 및 초기 burn-in 시간, 혼합률 제어 등 주의해야할 점들이 많다. 반면에, 근사화된 방법으로서 분석적으로 취급 가능할 정도의 근사화된 분포를 사용하여 적분을 해결하는 시도가 있다.

3. 변분 방법 (Variational Methods)

변분 베이지안 방법의 아이디어는, 사후 확률에 가깝게 근사하면서도 다루기 용이한 확률 분포를 찾는 것이다. 사후 확률에 근접시키는 원리는 다음과 같다.

$$\ln p(y_{1:T}) = \ln \int d\theta ds_{1:T} p(\theta) p(s_{1:T}, y_{1:T} \mid \theta)$$
$$\geq \int d\theta ds_{1:T} q(\theta, s_{1:T}) \ln \frac{p(\theta) p(s_{1:T}, y_{1:T} \mid \theta)}{q(\theta.s_{1:T})}$$

여기에서 부등호는 Jensen의 부등식에 의한 것이다. 즉, 적분이 용이한 분포 $q(\theta,s_{1:T})$를 이용하여, $\ln p(y_{1:T})$의 아래로 유계인 근사치를 계산하는 것이다. 변분 베이지안 방법은 유계 값을 최대화 하는 $q(\theta,s_{1:T})$를 찾는다. 그런 분포는 사후 분포와의 Kullback-Leibler 거리 $D(q(\theta,s_{1:T}) \| p(\theta|y_{1:T}))$ 를 최소화 하는 분포이다. 일반적으로 근사화 분포 $q(\theta,s_{1:T})$로 다음과 같이 결합도를 제거한 단순한 형태를 취한다.

$$p(\pi, A, M, \{\mu\}, \{\Lambda\}, s_{1:T} \mid y_{1:T})$$
$$\approx q(\pi) q(A) q(M) q(\{\mu\}) q(\{\Lambda\}) q(s_{1:T})$$

그림 1의 그래프 표현에서 부모 노드와 자식 노드간에 공액관계를 만족시키는 지수 분포를 사용하면, 닫힌 형태의 갱신식을 유도할 수 있다. 갱신 절차는 $q(\theta)$와 $q(s_{1:T})$를 서로 번갈아가며 최대화 하는 것으로 생각할 수 있다 [1]. 또한 베이지안 네트워크의 메시지 전송 기법과 유사한 변분 메시지 전송 형태의 일반화된 갱신 방법을 생각할 수 있다 [2, 7].

3.1. 갱신식

근사화 분포 $q(\theta,s_{1:T})$를 취하는 것은, 그림 1에서 θ와 관계된 노드들의 종속 관계를 배제시키는 것과 같다. 그러나, $q(s_{1:T})$의 체인은 종속 관계의 배제 없이 HMM의 추론 알고리즘을 통해 $p(s_{1:T} \mid y_{1:T}, \theta)$로의 정확한 일치가 가능하다. 갱신 절차는, 첫번째 단계에서는 $q(\theta)$를 고정시킨 후, $s_{1:T}$ 로의 메시지들을 취합하는 HMM forward-backward 기법을 수행한다. 두번째 단계에서, $q(s_{1:T})$가 고정되고, 자신의 메시지를 파라미터 노드들로 전파한다. 파라미터 노드들은 메시지를 취합하여 $q(\theta)$를 갱신한다.

$q(s_{1:T})$ 의 갱신은 [1] 과 유사하다. 단, 노드 $M^T s_t$ 에서 s_t로 가는 메시지는 다음의 식을 사용한다.

$$m_{M^T s_t \to s_t} = \sum_{k=1}^{K} \left(\exp\langle \ln M \rangle\right)_{s_t,k} \ln p(y_t \mid m(\mu_{s_t,k}), m(\Lambda_{s_t,k}))$$

$q(\theta)$의 갱신식은 다음과 같다 이 때, 〈·〉는 현재 갱신이 되는 분포를 제외한 다른 분포에 대한 평균 연산을 의미한다. $M^T s_t$ 에서 M로 가는 메시지는 다음과 같다.

$$m_{M^T s_t \to M_{sk}} = \ln p(y_t \mid m(\mu_{s_t,k}), m(\Lambda_{s_t,k}))$$

M의 분포 $M_{(s,:)} = Dir(m_{s1},...,m_{sK} \mid w_{s1}^{(M)},...,w_{sK}^{(M)})$ 는

$$w_{sk}^{(M)} = u_{sk}^{(M)} + \sum_{t=1}^{T} m_{M^T s_t \to M_{sk}}$$

과 같이 갱신한다. u_{sk}는 사전 파라미터이다.

혼합 노드로의 메시지는 다음과 같다.

$$m_{y_t \to \mu_{s,k}} = \left(\exp\langle \ln M \rangle\right)_{s,k} \times \left[\left\langle \Lambda_{s,k} \right\rangle y_t ; -\frac{1}{2} \mathrm{vec}\left(\left\langle \Lambda_{s,k} \right\rangle \right) \right]$$

$$m_{y_t \to \Lambda_{s,k}} = \left(\exp\langle \ln M \rangle\right)_{s,k}$$
$$\times \left[-\frac{1}{2} \mathrm{vec}\left(\left\langle \mu_{s,k} \mu_{s,k}^T \right\rangle - y_t \left\langle \mu_{s,k}^T \right\rangle - \left\langle \mu_{s,k} \right\rangle y_t^T + y_t y_t^T \right\}, \frac{1}{2} \right]$$

혼합 노드의 평균 $\mu_{s,k}$ 의 분포는 가우시안 분포로서, $q(\mu_{s,k} \mid \mu_{s,k}^*, \lambda_{s,k}^*)$의 정확률 행렬 및 평균 벡터를

$$\lambda_{s,k}^* = \lambda_0 + \sum_{t=1}^{T} m_{y_t \to \mu_{s,k}}(2),\ \mu_{s,k}^* = \lambda_{s,k}^{*\,-1} \left(\lambda_0 \mu_0 + m_{y_t \to \mu_{s,k}}(1) \right)$$

과 같이 갱신한다. 이 때, $m_{y_t \to \mu_{s,k}}(2)$ 은 이 메시지의 두번째 값을 사용하겠다는 의미이다. 정확도 행렬 Λ는 위샤트 분포로서, $q(\Lambda_{s,k} \mid \nu_{s,k}^*, \Lambda_{s,k}^*)$ 의 파라미터는

$$\Lambda^*_{s,k} = \left[\Lambda_0^{-1} + \sum_{t=1}^{T} m_{y_t \to \Lambda_{s,k}}(1) \right]^{-1}, \quad \nu^*_{s,k} = \nu_0 + \sum_{t=1}^{T} m_{y_t \to \Lambda_{s,k}}(2)$$

과 같이 갱신한다. π 와 A 분포의 갱신은 [1]의 경우와 동일하게 각각 $\langle s_t \rangle, \langle s_{t-1} s_t \rangle$를 이용하여 수행된다.

4. 실험

본 절에서는 인공 데이터에 대한 실험을 통해, 변분 베이지안 방법의 성능을 제시하려 한다. 그림 2와 같이 5개 은닉 상태를 지니는 HMM으로부터 길이 30인 샘플 100개를 생성하였다. 각 상태의 초기 확률은 동일하며, 각 상태에서 자신의 상태에 머무를 확률과 다음 상태로 전이할 확률은 같게 하였다. 화살표가 상태 전이를 나타내며 자신에 머무르는 것은 표현하지 않았다. 출력은 2차원 가우시안 분포에서 생성하였다. 이 데이터에 대해서 상태의 개수를 30개, 상태별 혼합 노드의 수는 3개로 고정후 학습을 진행 시켰다.

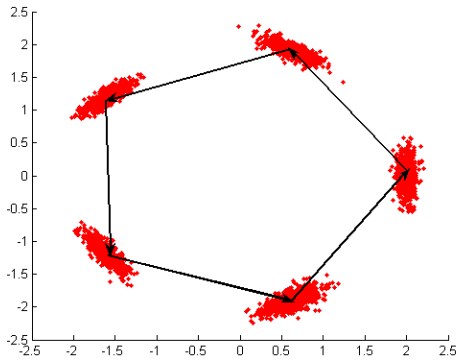

그림 2. 인공 데이터

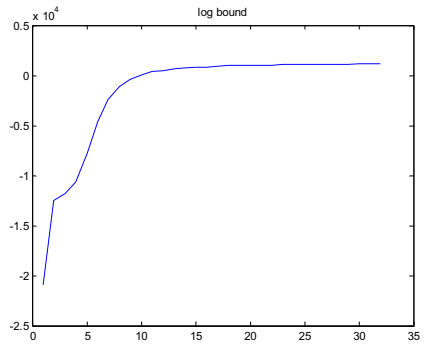

그림 3. 로그 바운드

그림 3은, 학습 데이터에 대한 $\ln p(\underline{y})$의 바운드를 나타낸다. 가로축은 파라미터의 갱신 횟수이다. 학습은 바운드의 상승이 임계치 이하가 될 때 중지하였다. 변분 메시지 전송에 의한 갱신은 바운드의 단조 증가성을 보장한다. 그림 4는, 학습 결과로서 (a)상태 전이, (b) 초기 상태, (c) 혼합 모델 확률 분포의 디리클렛 파라미터의 히튼 그림이다. 흰색 사각형은 디리클렛 파라미터의 크기를 나타낸다. 결과로 5개의 은닉 상태 및 혼합의 요인 수를 매우 유사하게 근사시킨다.

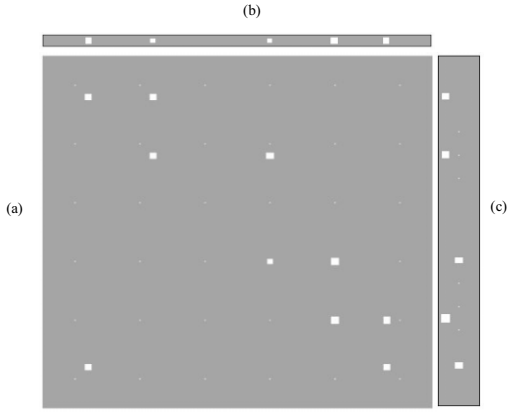

그림 4. 학습된 HMM의 파라미터 분포

5. 결론

가우시안 혼합 출력 노드를 지닌 HMM에 대한 베이지안 추론 및 학습법을 유도하였다. 결정론적 방법인 변분 베이지안 기법을 통해 최대우도 기법의 단점인 과학습 및 바이어스를 줄이고, 효과적인 모델 학습을 가능케 한다.

감사의 글

이 논문은 교육인적사업부의 BK21 사업과 과학기술부의 국가지정연구실 사업 (NRL)과 산업자원부에 의해 지원되었음.

6. 참고문헌

[1] M. J. Beal, Variational Algorithms for Approximate Bayesian Inference, PhD. Thesis, University College London, 2003.

[2] Z. Ghahramani and M. J. Beal, Propagation Algorithm for Variational Bayesian Learning, NIPS 13, 2001.

[3] Z. Ghahramani and G. E. Hinton, Variational Learning for Switching State-Space Models, Neural Computation, 12(4), pp. 831-864, 2000.

[4] V. Pavlovic, B. J. Frey and T. S. Huang, Variational Learning in Mixed-State Dynamic Graphical Models, UAI, 1999.

[5] L. R. Rabiner, A Tutorial in Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 77(2), pp. 257-286, 1999.

[6]S. Watanabe, Y. Minami, A. Nakamura and N. Ueda, Application of Variational Bayesian Approach to Speech Recognition, NIPS 15, 2003.

[7] J. Winn and C. M. Bishop, Variational Message Passing, Journal of Machine Learning Research, submitted. 2004.