워크북

교과목명 : 머신 러닝

차시명: 11차시

◈ 담당교수: **장 필 훈**

⊙ 세부목차

- 은닉마르코프모델의 특징
- hmm에서 최대가능도법
- 바움웰치
- 비터비
- hmm의 확장

학습에 앞서

■ 학습개요

순차데이터에서 가장 비중있는 모델로 은닉 마르코프 모델을 다루게 된다. 관찰변수를 야기한 은닉변수들을 추정하는 방법에 대해 식으로 자세히 다룬다. 그 과정에서 forward probability, backward probability등을 구체적으로 계산해내는 방법을 배우고 이를 이용해서 hmm을 EM알고리즘을 이용해 풀어내는 단계에 대해 자세히 익힌다. 바움-웰치 알고리즘과 비터비 알고리즘이 나오는데, 바움웰치 알고리즘을 우선 익히고, 그 단점을 극복할 수 있는 비터비 알고리즘을 배운다. 비터비알고리즘은 동적계획법을 따르는데, 그 아이디어에 대해서도 배운다.

hmm의 이론적인 면을 배운 뒤에는 실제로 사용했을 때의 단점을 배우고, 그 단점들을 극복하기 위한 hmm의 응용모델들에 관해 배운다. 입출력 hmm, factorial hmm등이 hmm의 응용에 속한다.

■ 학습목표

1	hmm의 구조와 가능도 함수를 식으로 이해한다.
2	forward probability와 backward probability를 이해하고 수식으로 구해내는 법을 익힌다.
3	hmm을 해결하는 알고리즘 중 하나인 바움-웰치 알고리즘을 이해한다.

4	바움-웰치 알고리즘의 약점을 극복할 수 있는 비터비 알고리즘을 이해한다.

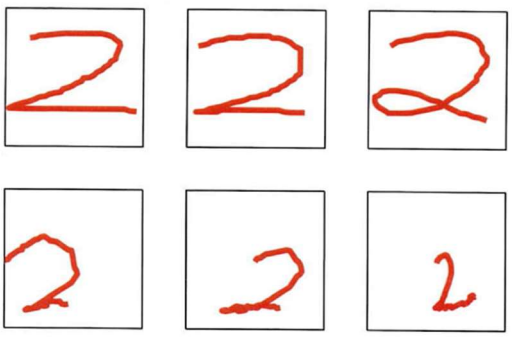
■ 주요용어

용어	해설
HMM	은닉마르코프 모델. 시스템이 관측변수와 은닉변수로만 이루어져 있다고 가정하고, 은닉변수가 마르코프 과정을 따른다. 딥러닝이 주로 쓰이기 이전에 음성인식분야의 주류였다.
transition probability	은닉상태의 전이확률. 마르코프조건에 따라 바로 이전상태가 다음상태를 결정하는데, 이때 각각의 가능한 상태끼리의 전이 가능한 확률. 이산변수의 경우 행렬모양으로 나타난다.
emission probability	특정시간에서 은닉상태→관찰상태로 가는 확률. 방사확률이라고 한다. 은닉마르코프 모델은 은닉상태만 연쇄형태를 가지고 관찰상태는 은닉상태에만 의존하는데, 이때 P(x|z)가 방사확률이다.
바움-웰치 알고리즘	은닉마르코프모형을 학습하는 방법. EM의 일종. 전이확률과 방출확률(emission probability)을 이용해서 매개변수를 조정하고, 다시 조정된 매개변수를 바탕으로 전이확률과 방출확률을 계산하는 식으로 이루어진다.
비터비 알고리즘	비터비경로(Viterbi path)를 찾기 위한 동적계획법(dynamic programming)알고리즘. 은닉마르코프 모형에서 관측된 변수들을 야기한 가장 가능성높은 은닉변수들의 sequence를 구해내는 알고리즘.

학습하기

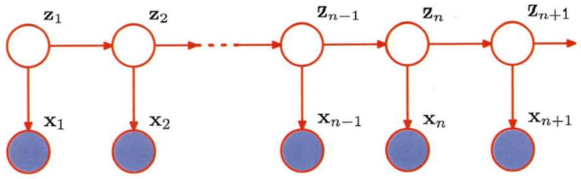
전시간에 이어, hmm에 관해 계속 공부하겠습니다.

hmm은 시간축의 뒤틀림에 강하다는 장점이 있습니다.



왼쪽 그림(Bishop. Fig.13.11)을 보면 시작점이 숫자 2의 시작점이고 그 시작점에 따라 hmm이 생성해내는 점들을 계속 그려나간 것입니다. 이 hmm은 숫자2를 그리도록 되어 있습니다. 그림에서 보듯이 어느점에서 시작하든지 2와 얼추 비슷하게 그려나가는 것을 관찰할 수 있습니다.

이제 hmm을 최대가능도법으로 학습시키는 방법을 살펴보겠습니다. 최대가능도법의 기본은 관측값을 이용해서 매개변수를 정하는 것입니다. 그런데 가능도함수가 매우 복잡한 형태일 경우 닫힌해를 구



하기 어려우므로 EM방법으로 푼다는 것을 앞서 배웠습니다.

hmm의 구조는 왼쪽 그림과 같고, Z가 hidden, X가 관찰인것도 저번시간에 배웠습니다. 최대가능도법을 사용하려면, transition probability, emission prob, forward prob, backward prob을 먼저 정의해야 합니다. 각각은 다음과 같습니다. (아래부터는 강의를 먼저 보시기를 강력히 권합니다. 식이 많아서 설명없이 보면 이해가 힘듭니다)

$Z_t \to Z_{t+1}$: transition probability A (Z_n의 분포를 w(column vector)로 나타냄)

$Z_t \to X_t$: emission probability B

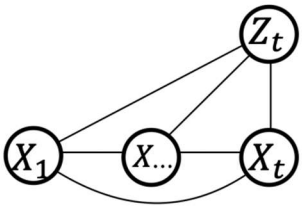
forward probability : $\alpha_t(i) = p_\theta(x_1, \dots, x_t, Z_t = i)$

backward probability : $\beta_t(i) = p_\theta(x_{t+1}, \dots, x_n | Z_t = i)$

이제 w, A, B를 이용해서 θ의 MLE를 구해보겠습니다.

모든 관측값이 있을 때, latent states의 가능도는 $p_\theta(Z_1, \dots, Z_n | x)$입니다.

forward probability는 그림으로 나타내면 오른쪽과 같습니다.



식으로 나타내면,

$$\alpha_t(i) = p_\theta(x_1, \dots, x_t, Z_t = i)$$

$$\begin{aligned}\alpha_1(i) &= p_\theta(x_1, Z_1 = i) \\ &= p_\theta(Z_1 = i) \times p_\theta(x_1 | Z_1 = i) \\ &= w(i) B(i, x_1)\end{aligned}$$

$$\begin{aligned}\alpha_2(i) &= p_\theta(x_1, x_2, Z_2 = i) \\ &= \sum_j p_\theta(x_1, x_2, Z_1 = j, Z_2 = i) \\ &= \sum_j p_\theta(x_1, Z_1 = j) \times p_\theta(Z_2 | x_1, Z_1 = j) \times p_\theta(x_2 | x_1, Z_1 = j, Z_2 = i) \\ &= \sum_j \alpha_1(j) A(j, i) B(i, x_2)\end{aligned}$$

$$\begin{aligned}\alpha_{t+1}(i) &= p_\theta(x_1, \cdots, x_{t+1}, Z_{t+1} = i) \\ &= \sum_j p_\theta(x_1, \cdots, x_{t+1}, Z_t = j, Z_{t+1} = i) \\ &= \sum_j p_\theta(x_1, \cdots, x_t, Z_t = j) \times p_\theta(Z_{t+1} | x_1, \cdots, x_{t+1}, Z_t = j) \\ &\qquad \times p_\theta(x_{t+1} | x_1, \cdots, x_t, Z_t = j, Z_{t+1} = i) \\ &= \sum_j \alpha_1(j) A(j, i) B(i, x_2)\end{aligned}$$

다음은 backward probability계산입니다.

$$\begin{aligned}\beta_{n-1}(i) &= p_\theta(x_n|Z_{n-1}=i)\\ &= \sum_j p_\theta(x_n, Z_n=j|Z_{n-1}=i)\\ &= \sum_j p_\theta(Z_n=j|Z_{n-1}=i)\times p_\theta(x_n|Z_n=j, Z_{n-1}=i)\\ &= A(i,j)B(j,x_n)\\ &= A(i,j)B(j,x_n)\beta_n(j) \qquad (\beta_n(*)\equiv 1)\end{aligned}$$

$\beta_n(*)$를 1로 정의한것에 주의하세요. 개념적인 정의입니다. (어떻게 보면 당연한 가정입니다)

이제 전체에 대한 가능도 함수를 적으면, .

$$p(\mathbf{x}_1,\dots,\mathbf{x}_N,\mathbf{z}_1,\dots,\mathbf{z}_N) = \mathrm{p}(\mathbf{z}_1)\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1})\right]\prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

(로그가능도)

$$\begin{aligned}&\log\left[w(Z_1)\prod_{t=1}^{n-1}A(Z_t,Z_{t+1})\prod_{t=1}^{n}B(Z_t,x_t)\right]\\ &= \log w(Z_1) + \sum_{t=1}^{n-1}\log A(Z_t,Z_{t+1}) + \sum_{t=1}^{n}\log B(Z_t,x_t)\end{aligned}$$

이제 이것을 최대화하면 됩니다.

바움-웰치 방법은 새로운 변수 감마를 정의합니다.

$$\gamma_t(i,j) = p_\theta(Z_t=i, Z_{t+1}=j|\mathbf{x}),$$
$$\gamma_t(i) = p_\theta(Z_t=i|\mathbf{x}) = \sum_j \gamma_t(i,j)$$

로그가능도 함수를 이 감마에 대해 나타내면 다음과 같습니다.

$$\begin{aligned}&\mathbb{E}_{(\mathbf{Z}|\mathbf{x},\theta)}\log p(\mathbf{Z},\mathbf{x}|\theta)\\ &= \mathbb{E}_{(\mathbf{Z}|\mathbf{x},\theta)}\left[\log w(Z_1) + \sum_{t=1}^{n-1}\log A(Z_t,Z_{t+1}) + \sum_{t=1}^{n}\log B(Z_t,x_t)\right]\end{aligned}$$

$$= \sum_i^{m_z}\gamma_1(i)\log w(i) + \sum_{t=1}^{n-1}\sum_{i,j=1}^{m_z}\gamma_t(i,j)\log A(i,j) + \sum_{t=1}^{n}\sum_{i=1}^{m_z}\gamma_t(i)\log B(i,x_t)$$

$$= \sum_i^{m_z}\gamma_1(i)\log w(i) + \sum_{i,j=1}^{m_z}\sum_{t=1}^{n-1}\gamma_t(i,j)\log A(i,j) + \sum_{i=1}^{m_z}\sum_{t=1}^{n}\gamma_t(i)\log B(i,x_t)$$

$$= \sum_{i}^{m_z} \gamma_1(i) \log w(i) + \sum_{i,j=1}^{m_z} \sum_{t=1}^{n-1} \gamma_t(i,j) \log A(i,j) + \sum_{i=1}^{m_z} \sum_{t=1}^{n} \gamma_t(i) \log B(i, x_t)$$

$$= \sum_{i}^{m_z} \gamma_1(i) \log w(i) + \sum_{i}^{m_z} \left[\sum_{j=1}^{m_z} \left[\sum_{t=1}^{n-1} \gamma_t(i,j) \right] \log A(i,j) \right]$$

$$+ \sum_{i=1}^{m_z} \sum_{l=1}^{m_x} \left(\sum_{\text{at } x_t = l} \gamma_t(i) \right) \log B(i,l) \qquad \because x_t \in \{m_x\}$$

식을 천천히 보면 크게 어려운것은 없습니다만, 이해가 안가시면 강의를 참고하셔도 좋고, 제가 참고했던Feng Liang, Stat542_W7_HMM강의를 참고해도 됩니다.(Youtube에 공개되어 있습니다)

이제 위의 식을 최대화하기만 하면 됩니다(M step)

위의 식은 간단히 추상화하면 다음과 같이 나타낼 수 있고,

$$a_1 \log b_1 + a_2 \log b_2 + \cdots + a_m \log b_m$$

$$\text{where } a_i > 0,\ b_i > 0,\ \textstyle\sum_i a_i = 1,\ \sum_i b_i = 1$$

해당 함수는 $a_i = b_i$일 때 최댓값을 가짐을 보일 수 있습니다.(강의에 보였습니다. 참고하세요)

그에 따라 위 바움웰치에서 구했던 식을 최대화하는 조건을 구할 수 잇습니다.

첫번째 항은 $\gamma_1(i) = w(i)$, 두번째항은

$$A(i,j) = \frac{\sum_{t=1}^{n-1} \gamma_t(i,j)}{\sum_{j'} \sum_{t=1}^{n-1} \gamma_t(i,j')}, \qquad i,j = 1, \ldots, m_z$$

세번째 항은,

$$B(i,l) = \frac{\sum_{\text{at } x_t = l} \gamma_t(i)}{\sum_t \gamma_t(i)}$$

$\gamma_t(i,j)$는 A, B, alpha, beta로 다시 나타낼 수 있습니다.

$$p_\theta(Z_t = i, Z_{t+1} = j | \mathbf{x})$$

$$\propto p_\theta(x_1, \ldots, x_t, Z_t = i, Z_{t+1} = j, x_{t+1}, \ldots, x_n)$$

$$= p_\theta(x_1, \ldots, x_t, Z_t = i) \times p_\theta(Z_{t+1} = j | Z_t = i)$$

$$\times p_\theta(x_{t+1} | Z_{t+1} = j) \times p_\theta(x_{t+2}, \ldots, x_n | Z_{t+1} = j)$$

$$= \alpha_t(i) A(i,j) B(j, x_{t+1}) \beta_{t+1}(j)$$

Z_t 각각에 대해 가장 optimal한 값을 선택하는 방식은 각각의 t에 대해 최적의 상태를 선택하므로 sequence가 유효하지 않을 수 있다는 문제가 있습니다. transition prob=0인데 뒤이은 상태가 '불가능한' 상태로 선택될 수 있다는 뜻입니다.

따라서 가능한 sequence중에 가장 그럴듯한 것을 찾아야 합니다. 그것을 찾는 알고리즘이 비터비

알고리즘입니다. 비터비방법은 델타라는 새로운 변수를 정의하고 다음을 찾습니다.

$$Z^* = \arg\max_{i_1,\cdots,i_n} p_\theta(Z_1 = i_1, \cdots, Z_n = i_n|\mathbf{x})$$

델타는 다음과 같이 정의되고, δ_1부터 차례로 구할 수 있습니다.

$$\delta_t(i) = \max_{j_1,\cdots,j_{t-1}} p_\theta(Z_1 = j_1, \ldots, Z_{t-1} = j_{t-1}, Z_t = i, x_1, \cdots, x_t)$$

$$\delta_1(i) = p_\theta(Z_1 = i, x_1) = w(i)B(i, x_1)$$

$$\delta_{t+1}(i) = \max_{j_1,\cdots,j_{t-1},j} p_\theta(Z_1 = j_1, \cdots, Z_{t-1} = j_{t-1}, Z_t = j, Z_{t+1} = i, x_1, \cdots, x_{t+1})$$

$$= \max_{j_1,\cdots,j_{t-1},j} [p_\theta(Z_1 = j_1, \cdots, Z_{t-1} = j_{t-1}, Z_t = j, x_1, \cdots, x_t) \times p_\theta(Z_{t+1} = i|Z_t = j) \times p_\theta(x_{t+1}|Z_{t+1} = i)]$$

$$= \left[\max_j \delta_t(j)A(j,i)\right] B(i, x_{t+1})$$

Z와 δ는 다음의 관계를 갖습니다.

$$Z^*_{n-1} = \arg\max_i [\delta_{n-1}(i)A(i, j^*_n)]$$

이를 이용해서 Z^*를 구할 수 있습니다.

hmm도 여러가지 단점을 가집니다. 가장 큰 단점은, 주어진 상태를 얼마나 유지하는지 나타내기에 부적합합니다. 전이확률을 고려해보면, 제자리에 머무르기 위해 확률이 계속 곱해져야 하므로 기하급수적으로 감쇠하기 때문입니다. 그리고 시간적으로 거리가 먼 관측변수들간의 상관관계를 잘 나타내기도 쉽지 않습니다.

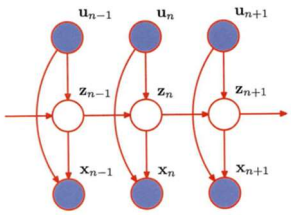
이를 해결하기 위해 autoregressive hmm, input-output hmm, factorial hmm등이 있습니다. 각각의 내용은 강의에서 간단히 다루었습니다.

	연습문제

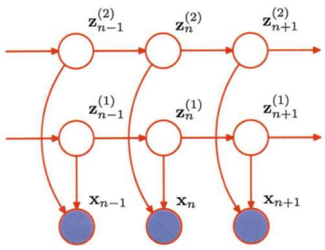
1. 바로 이전 은닉상태에만 의존하는 hmm은 생성모델로 쓸 수 없다.
    a. O
    b. 성능이 좋지 않을 뿐 쓰는데는 문제가 없다. 어느정도 결과를 보여준다.
2. hmm에서 매개변수를 구하기 위해 최대가능도법을 쓰면 완전히 닫힌 해를 쉽게 얻을 수 있다.
    a. X
    b. 혼합모델의 경우와 동일하게, 로그 안에 합산항이 존재하기 때문에 완전히 닫힌 해를 계

산해 내는 것이 매우 어렵다.

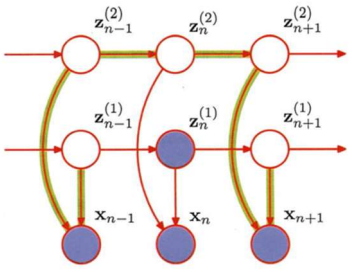
3. 바움웰치 방법으로 Z_n을 각각 추정해내면, sequence가 유효하지 않을 수 있다.
   a. O
   b. transition probability가 0인데도 전후 상태가 연속될 수 있다.
4. 유효한 최적의 sequence를 찾아내기 위해 쓰는 비터비 알고리즘은 dynamic programming방법에 기반한다.
   a. O
   b. 연쇄적인 관계를 이용하여 차례로 모두 구해내는 방법.
5. 기본적인 hmm을 사용하면 거리가 먼 관측변수들간의 관계를 모델링해내기가 어렵다.
   a. O
   b. 바로 이전단계에만 의존하기 때문에 거리가 멀면 관계를 알기 어렵다.
6. 아래 그림과 같은 입출력 hmm은 마르코프성질($\mathrm{z_{n-1}} \perp\!\!\!\perp z_{n+1} \mid z_n$)을 만족한다.



   a. O
   b. z_n이 관측되면 $\mathrm{z_{n-1}}$과 z_{n+1}은 독립이게 된다. $\mathrm{z_n}$이 head-to-tail 노드이기 때문.
7. 아래 그림과 같은 factorial hmm은 마르코프성질($\mathrm{z_{n-1}} \perp\!\!\!\perp z_{n+1} \mid z_n$)을 만족한다.



   a. X
   b. 아래 녹색 경로로 열려있다.



정리하기

1. hmm은 시간축의 뒤틀림에 강하지만, 생성모델로서는 적당하지 않다.
    a. 바로 이전 상태에만 의존하는것이 기본이어서 생성품질이 뛰어난 편이 아니다.
2. hmm에서 forward probability는 다음과 같이 계산된다.

$$\alpha_{t+1}(i) = \sum_j \alpha_t(j) A(j,i) B(i, x_{t+1})$$

3. hmm에서 backward probability는 다음과 같이 계산된다.

$$\beta_{n-1}(i) = A(i,j) B(j, x_n) \beta_n(j)$$

4. hmm에서는 EM알고리즘을 사용해서 매개변수를 추정해낸다.(바움-웰치 알고리즘)
5. 바움-웰치로 추정해 낸 은닉상태는, transition을 고려하지 않기 때문에, 최적의 sequnce를 추정해 내기 위해서는 비터비 알고리즘을 쓴다.
    a. 비터비 알고리즘은 recursion을 이용한다.
6. hmm은 상태유지의 시간분포를 나타내는 데 부적합하다.
    a. 확률의 power(승)으로 값이 나타나기 때문에 기하급수적으로 감소한다.
7. hmm은 거리가 먼 관측변수들간 상관관계를 잘 잡아내지 못한다.
8. 단점을 보완하기 위해 여러가지 hmm의 변형들이 존재한다.
    a. 자기회귀적 은닉마르코프 모델 : x_{n-2}까지 의존성을 더한 모델
    b. 입출력 hmm : 관측변수가 하나 더 있는 모델
    c. factorial hmm : 출력 하나하나의 조합으로 최종 출력을 나타내는 모델

참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

다음 차시 예고

- 순차데이터
    - LDS
    - 칼만필터
- 모델조합
    - bagging
    - adaboost
    - gradient boost