隠れマルコフモデル

正田 備也

masada@rikkyo.ac.jp

Contents

マルコフモデル

- $\boldsymbol{z} = (z_1, \ldots, z_T)$ という確率変数の列があるとする
- z_t は、状態の集合 $\mathcal{S} = \{s_1, \ldots, s_K\}$ の要素を値として取る

  例. 天候の集合 $\mathcal{S} = \{s_{\mathsf{sun}}, s_{\mathsf{cloud}}, s_{\mathsf{rain}}\}$
- 以下の単純マルコフ性の仮定をおく

$$p(z_i = s | z_1, \ldots, z_{i-1}) = p(z_i = s | z_{i-1}) \text{ for all } s \in \mathcal{S} \tag{1}$$

- s_k から s_l に遷移する確率 $p(z_t = s_l | z_{t-1} = s_k)$ は、全ての t で等しいと仮定（斉時性の仮定）し、この確率を A_{s_k,s_l} と書く
  - $\sum_{l=1}^{K} A_{s_k,s_l} = 1$ が、すべての k で成り立つ
- 便宜的に初期状態を確率変数 z_0 で表し、その値を s_0 とする
  - $A_{s_k,s_0} = 0$ が、すべての k について成り立つ

遷移行列

- 遷移確率をまとめて、遷移行列 A として書く

例. 天候の状態の集合 $S = \{s_0, s_{\mathsf{sun}}, s_{\mathsf{cloud}}, s_{\mathsf{rain}}\}$

$$A = \begin{array}{c|cccc} & s_0 & s_{\mathsf{sun}} & s_{\mathsf{cloud}} & s_{\mathsf{rain}} \\ \hline s_0 & 0 & 0.33 & 0.33 & 0.33 \\ s_{\mathsf{sun}} & 0 & 0.8 & 0.1 & 0.1 \\ s_{\mathsf{cloud}} & 0 & 0.2 & 0.6 & 0.2 \\ s_{\mathsf{rain}} & 0 & 0.1 & 0.2 & 0.7 \end{array} \tag{2}$$

cf. http://cs229.stanford.edu/section/cs229-hmm.pdf

状態列の尤度

- 状態の列 $\boldsymbol{z}_{1:T} \equiv (z_1, \ldots, z_T)$ の尤度を $p(\boldsymbol{z}_{1:T}; A)$ と書く

$$
\begin{aligned}
& p(\boldsymbol{z}_{1:T}; A) \\
& = p(z_0, z_1, z_2, \ldots, z_T; A) \\
& = p(z_T | z_{T-1}, \ldots, z_1; A) \cdots p(z_3 | z_2, z_1, z_0; A) p(z_2 | z_1, z_0; A) p(z_1 | z_0; A) \\
& = p(z_T | z_{T-1}; A) \cdots p(z_2 | z_1; A) p(z_1 | z_0; A) \\
& = \prod_{t=1}^{T} p(z_t | z_{t-1}; A) = \prod_{t=1}^{T} A_{z_{t-1}, z_t} \qquad (3)
\end{aligned}
$$

問. どこで単純マルコフ性の仮定を使っているか？

マルコフモデルの最尤推定

$\ln p(\boldsymbol{z}_{1:T}; A)$ を最大化することで A を推定 estimate する。

$$\ln p(\boldsymbol{z}_{1:T}; A) = \sum_{t=1}^{T} \ln A_{z_{t-1}, z_t} = \sum_{t=1}^{T} \sum_{l=1}^{K} \sum_{k=1}^{K} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k) \ln A_{s_l, s_k} \tag{4}$$

$\mathbb{1}(\cdot)$ は、括弧内の命題が真のとき 1、偽のとき 0、という意味だとする。

$$\mathcal{L}(A, \{\lambda_k\}) = \sum_{t=1}^{T} \sum_{l=1}^{K} \sum_{k=1}^{K} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k) \ln A_{s_l, s_k} + \sum_{l=1}^{K} \lambda_l \left(1 - \sum_{k=1}^{K} A_{s_l, s_k}\right) \tag{5}$$

$$\frac{\partial \mathcal{L}(A, \{\lambda_k\})}{\partial A_{s_l, s_k}} = \frac{\sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)}{A_{s_l, s_k}} - \lambda_l \tag{6}$$

$\frac{\partial \mathcal{L}(A, \{\lambda_k\})}{\partial A_{s_l, s_k}} = 0$ より $A_{s_l, s_k} \propto \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)$ で、$\sum_{k=1}^{K} A_{s_l, s_k} = 1$ だから

$$\therefore A_{s_l, s_k} = \frac{\sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)}{\sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)} = \frac{\sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)}{\sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l)} \tag{7}$$

Contents

例. アイスクリームの売上で温暖化を研究

- あなたは西暦 2799 年の世界に生きており、地球温暖化の研究をしている
- 歴史的な資料の中から、2020 年夏の毎日のアイスクリームの売上の記録が見つかった
- この記録を元に、2020 年夏の毎日の天候を推定したい
  - 天候の観測データは失われてしまっている！
- ただし、アイスクリームの売上は天候だけに依存すると仮定
  - https://web.stanford.edu/~jurafsky/slp3/A.pdf

隠れマルコフモデル HMM; hidden Markov model

- 状態 z_t は観測できず、各状態が生成する結果 x_t だけが観測できるとする
  - このとき、状態を表す確率変数 z_t は潜在変数 latent variable となる
- 隠れ状態の列 $\boldsymbol{z}_{1:T} = (z_1, \ldots, z_T)$ は、上述のとおり、単純マルコフ性と斉時性を持つと仮定する
- 時点 t での観測結果を、確率変数 x_t で表す
- 観測データ x_t について、以下のような独立性の仮定をおく

$$p(x_t = v_w | \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(x_t = v_w | z_t = s_k) \tag{8}$$

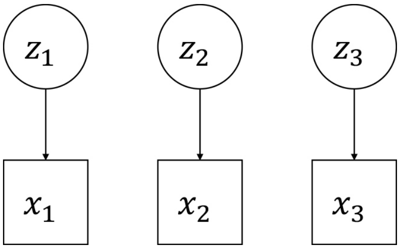
  - つまり、時点 t の観測値は、同じ時点の隠れ状態だけに依存する

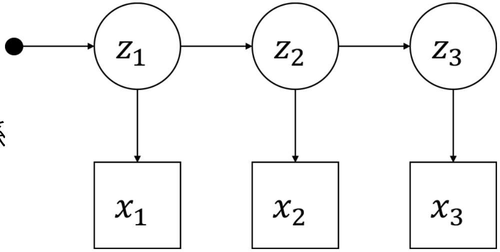
観測データがカテゴリカルデータの場合

- ここでは、観測データはカテゴリカルデータだとする
  - 観測データが連続値で、$p(x_t|z_t)$ が、例えば正規分布の場合、以下の議論がどうなるか、考えてみよう
- 隠れ状態 s_k がであるときの、アイテム v_w の出現確率 $p(x_t = v_w|z_t = s_k)$ を、B_{s_k,v_w} と書くことにする
  - $(B_{s_k,v_1}, \ldots, B_{s_k,v_W})$ は、状態 s_k に対応するカテゴリカル分布の、パラメータである
  - $\sum_{w=1}^{W} B_{s_k,v_w} = 1$ が成り立つ
- 隠れマルコフモデルにおけるパラメータ推定では、A と B を推定することになる

混合分布モデルと隠れマルコフモデル

- 混合分布モデル
  - 隠れ変数 z_i は独立





- 隠れマルコフモデル
  - 隠れ変数 z_t の間に依存関係
  - 線形で一方向的な依存関係

同時分布

- 隠れマルコフモデルでの観測変数と隠れ変数の同時分布は

$$
\begin{aligned}
p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B) &= p(\boldsymbol{x}_{1:T} | \boldsymbol{z}_{1:T}; B) p(\boldsymbol{z}_{1:T}; A) \\
&= \prod_{t=1}^{T} p(x_t | z_t; B) \times \prod_{t=1}^{T} p(z_t | z_{t-1}; A) \\
&= \prod_{t=1}^{T} B_{z_t, x_t} \times \prod_{t=1}^{T} A_{z_{t-1}, z_t}
\end{aligned} \tag{9}
$$

  - 状態列の確率 $p(\boldsymbol{z}_{1:T})$ の部分は、式 (3) と同じ
  - ただし、隠れマルコフモデルでは z_t は潜在変数

観測データの尤度

- 観測データの尤度 $p(\boldsymbol{x}; A, B)$ は

$$
\begin{aligned}
p(\boldsymbol{x}; A, B) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}; A, B) \\
&= \sum_{z_1 \in \mathcal{S}} \cdots \sum_{z_T \in \mathcal{S}} \left(\prod_{t=1}^{T} p(x_t | z_t; B) \right) \left(\prod_{t=1}^{T} p(z_t | z_{t-1}; A) \right) \\
&= \sum_{z_1 \in \mathcal{S}} \cdots \sum_{z_T \in \mathcal{S}} \left(\prod_{t=1}^{T} B_{z_t, x_t} \right) \left(\prod_{t=1}^{T} A_{z_{t-1}, z_t} \right) \qquad (10)
\end{aligned}
$$

  - だが、この式をそのまま使って計算すると、計算量が $O(|\mathcal{S}|^T) = O(K^T)$ であるため、非現実的な計算時間になる

Contents

Forward procedure

- HMM のデータ尤度 $p(\boldsymbol{x}; A, B)$ は、動的計画法の一種である forward procedure によって、効率的に計算できる
  cf. https://web.stanford.edu/~jurafsky/slp3/A.pdf
- 時点 t までのデータ列 $(x_1, \ldots, x_t)$ の確率と、時点 t での隠れ状態 z_t が s_k である確率の同時確率を、$\alpha_t(k)$ とおく。つまり

$$\alpha_t(k) \equiv p(x_1, \ldots, x_t, z_t = s_k; A, B) \tag{11}$$

- すると、データ尤度を以下のように表すことができる

$$p(\boldsymbol{x}; A, B) = \sum_{k=1}^{K} p(\boldsymbol{x}_{1:T}, z_T = s_k; A, B) = \sum_{k=1}^{K} \alpha_T(k) \tag{12}$$

$$
\begin{aligned}
\alpha_1(k) &= p(x_1, z_1 = s_k) = p(z_0 = s_0)p(z_1 = s_k|z_0 = s_0)p(x_1|z_1 = s_k) = A_{s_0,s_k}B_{s_k,x_1} && (13)\\
\alpha_2(k) &= p(x_1, x_2, z_2 = s_k) = \sum_{z_1 \in \mathcal{S}} p(x_1, x_2, z_1, z_2 = s_k)\\
&= \sum_{z_1 \in \mathcal{S}} p(x_1, z_1)p(z_2 = s_k|x_1, z_1)p(x_2|z_2 = s_k, x_1, z_1) && (14)\\
&= \sum_{z_1 \in \mathcal{S}} p(x_1, z_1)p(z_2 = s_k|z_1)p(x_2|z_2 = s_k) = \sum_{l=1}^{K} \alpha_1(l)A_{s_l,s_k}B_{s_k,x_2} && (15)\\
\alpha_3(k) &= p(x_1, x_2, x_3, z_3 = s_k) = \sum_{z_2 \in \mathcal{S}} p(x_1, x_2, x_3, z_2, z_3 = s_k)\\
&= \sum_{z_2 \in \mathcal{S}} p(x_1, x_2, z_2)p(z_3 = s_k|x_1, x_2, z_2)p(x_3|z_3 = s_k, x_1, x_2, z_2) && (16)\\
&= \sum_{z_2 \in \mathcal{S}} p(x_1, x_2, z_2)p(z_3 = s_k|z_2)p(x_3|z_3 = s_k) = \sum_{l=1}^{K} \alpha_2(l)A_{s_l,s_k}B_{s_k,x_3} && (17)
\end{aligned}
$$

式 (14) から (15)、式 (16) から (17) は、自明でない。→ 条件付き独立性

$$
\begin{aligned}
\alpha_t(k) &= p(\boldsymbol{x}_{1:t}, z_t = s_k) \\
&= \sum_{z_{t-1} \in \mathcal{S}} p(\boldsymbol{x}_{1:t-1}, x_t, z_{t-1}, z_t = s_k) \\
&= \sum_{z_{t-1} \in \mathcal{S}} p(\boldsymbol{x}_{1:t-1}, z_{t-1}) p(z_t = s_k | \boldsymbol{x}_{1:t-1}, z_{t-1}) p(x_t | z_t = s_k, \boldsymbol{x}_{1:t-1}, z_{t-1}) && (18) \\
&= \sum_{z_{t-1} \in \mathcal{S}} p(\boldsymbol{x}_{1:t-1}, z_{t-1}) p(z_t = s_k | z_{t-1}) p(x_t | z_t = s_k) = \sum_{l=1}^{K} \alpha_{t-1}(l) A_{s_l, s_k} B_{s_k, x_t} && (19)
\end{aligned}
$$

式 (18) から (19) は、自明でない。下記の条件付き独立性を示す必要あり。

z_{t-1} が所与のもとで z_t と $\boldsymbol{x}_{1:t-1}$ は条件付き独立。

z_t が所与のもとで x_t と $\boldsymbol{x}_{1:t-1}, z_{t-1}$ は条件付き独立。

条件付き独立性 conditional independence

- C が与えられたときに A と B が独立である、つまり、$p(A,B|C)=p(A|C)p(B|C)$ が成り立つとき、A と B は C が所与のもとで条件付き独立である、と言う
- 条件付き確率の定義より、$p(A,B|C)=p(A|B,C)p(B|C)$ はいつでも成り立つ。よって

$$A \text{ と } B \text{ は } C \text{ が所与のもとで条件付き独立である}$$
$$\Leftrightarrow p(A|B,C)=p(A|C) \tag{20}$$

注. $p(A,B|C)=p(A|C)p(B|C)$ と $p(A,B)=p(A)p(B)$ は別物

cf. https://www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php

問9.1

- $p(A,B,C|D)=p(A|D)p(B,C|D)$、つまり、
  D が所与のもとで A と B,C が条件付き独立ならば、
  $p(A,B|D)=p(A|D)p(B|D)$、つまり、
  D が所与のもとで A と B が条件付き独立になることを、
  示せ。

答え $p(A,B,C|D)=p(A|D)p(B,C|D)$ の両辺を C について周辺化すると、左辺は $\sum_C p(A,B,C|D)=p(A,B|D)$ となり、右辺は $\sum_C p(A|D)p(B,C|D)=p(A|D)p(B|D)$ となる。
よって、$p(A,B|D)=p(A|D)p(B|D)$ が言える。

$p(z_t|z_{t-1}, \boldsymbol{x}_{1:t-1}) = p(z_t|z_{t-1})$ を示す。

$$p(z_t|z_{t-1}, \boldsymbol{x}_{1:t-1}) = \frac{p(z_t, z_{t-1}, \boldsymbol{x}_{1:t-1})}{p(z_{t-1}, \boldsymbol{x}_{1:t-1})} \tag{21}$$

$$p(z_{t-1}, \boldsymbol{x}_{1:t-1}) = \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u) \tag{22}$$

$$\begin{aligned} p(z_t, z_{t-1}, \boldsymbol{x}_{1:t-1}) &= \sum_{\boldsymbol{z}_{1:t-2}} p(z_t|z_{t-1}) \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u) \\ &= p(z_t|z_{t-1}) \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u) \end{aligned} \tag{23}$$

$$\therefore p(z_t|z_{t-1}, \boldsymbol{x}_{1:t-1}) = \frac{p(z_t|z_{t-1}) \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u)}{\sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u)} = p(z_t|z_{t-1}) \tag{24}$$

z_{t-1} が所与のもとで z_t と $\boldsymbol{x}_{1:t-1}$ は条件付き独立であることが言えた。

$p(x_t|z_t, z_{t-1}, \boldsymbol{x}_{1:t-1}) = p(x_t|z_t)$ を示す。

$$p(x_t|z_t, z_{t-1}, \boldsymbol{x}_{1:t-1}) = \frac{p(x_t, z_t, z_{t-1}, \boldsymbol{x}_{1:t-1})}{p(z_t, z_{t-1}, \boldsymbol{x}_{1:t-1})} \tag{25}$$

式 (23) より

$$p(z_t, z_{t-1}, \boldsymbol{x}_{1:t-1}) = p(z_t|z_{t-1}) \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u)$$

$$\begin{aligned} p(x_t, z_t, z_{t-1}, \boldsymbol{x}_{1:t-1}) &= \sum_{\boldsymbol{z}_{1:t-2}} p(x_t|z_t) p(z_t|z_{t-1}) \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u) \\ &= p(x_t|z_t) p(z_t|z_{t-1}) \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u) \end{aligned} \tag{26}$$

$$\therefore p(x_t|z_t, z_{t-1}, \boldsymbol{x}_{1:t-1}) = \frac{p(x_t|z_t) p(z_t|z_{t-1}) \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u)}{p(z_t|z_{t-1}) \sum_{\boldsymbol{z}_{1:t-2}} \prod_{u=1}^{t-1} p(z_u|z_{u-1}) p(x_u|z_u)} = p(x_t|z_t) \tag{27}$$

z_t が所与のとき x_t と $\boldsymbol{x}_{1:t-1}, z_{t-1}$ は条件付き独立であることが言えた。

Contents

観測データから隠れ状態を求める (decoding)

- 観測データ $\boldsymbol{x}_{1:T}$ に対して、一番ありえそう most likely な隠れ状態の列 $\boldsymbol{z}_{1:T}$ を見つけたい
  - 潜在変数を含む確率モデルにおいて、与えられた観測データに対して一番あり得そうな潜在変数の値を求めることを decoding と呼ぶ
- そこで、$p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}; A, B)$ を最大にする各 z_t の値を求める
  - パラメータ A, B の値は分かっていると仮定する
- 下記が成り立つことに注意

$$\arg\max_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}; A, B) = \arg\max_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A, B)$$

- 以下、$\arg\max_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A, B)$ を求める

Viterbi アルゴリズム

- $\arg\max_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A, B)$ を求めるためのアルゴリズム
  - これも動的計画法
- 実は forward procedure とほとんど同じ計算をする
- 式 (13) 以降に現れる、隠れ状態に関する和 $\sum_{z_{t-1}\in\mathcal{S}}$ を、$\max_{z_{t-1}\in\mathcal{S}}$ で置き換えて、計算を進めればいいだけ
- ただし、そのとき、同時に $\arg\max_{z_{t-1}\in\mathcal{S}}$ も記録しておく
- $t = T$ まで計算が終わったら、途中で記録しておいた $\arg\max_{z_{t-1}\in\mathcal{S}}$ を逆にたどることで、隠れ状態の列が得られる

$$
\begin{aligned}
v_1(k) &= p(x_1, z_1 = s_k) = p(z_0 = s_0)p(z_1 = s_k|z_0 = s_0)p(x_1|z_1 = s_k) = A_{s_0,s_k}B_{s_k,x_1} \\
v_2(k) &= \max_{z_1 \in \mathcal{S}} p(x_1, x_2, z_1, z_2 = s_k) \\
&= \max_{z_1 \in \mathcal{S}} p(x_1, z_1)p(z_2 = s_k|z_1)p(x_2|z_2 = s_k) = \max_{l=1}^{K} v_1(l)A_{s_l,s_k}B_{s_k,x_2} && (28) \\
c_2(k) &= \arg\max_{l=1}^{K} v_1(l)A_{s_l,s_k}B_{s_k,x_2} && (29) \\
v_3(k) &= \max_{z_2 \in \mathcal{S}} \left(\max_{z_1 \in \mathcal{S}} p(x_1, x_2, z_1, z_2 = s_k) \right) p(z_3 = s_k|z_2)p(x_3|z_3 = s_k) \\
&= \max_{l=1}^{K} v_2(l)A_{s_l,s_k}B_{s_k,x_3} && (30) \\
c_3(k) &= \arg\max_{l=1}^{K} v_2(l)A_{s_l,s_k}B_{s_k,x_3} && (31) \\
\text{cf. } \alpha_3(k) &= \sum_{z_2 \in \mathcal{S}} \left(\sum_{z_1 \in \mathcal{S}} p(x_1, x_2, z_1, z_2) \right) p(z_3 = s_k|z_2)p(x_3|z_3 = s_k) && (32)
\end{aligned}
$$

周辺化して z_{t-1} を消去するための確率の和の計算を、和をとられている確率のうちの最大値を選びとる計算に置き換える

$c_t(k)$ は、$z_t = s_k$ とするとき、一時点前の状態として何を選べばよいかを、表す

Viterbi アルゴリズム

- initialization
  - $v_1(k) = A_{s_0,s_k} B_{s_k,x_1}$
  - $c_1(k) = s_0$
- recursion
  - $v_t(k) = \max_l v_{t-1}(l) A_{s_l,s_k} B_{s_k,x_t}$
  - $c_t(k) = \arg\max_l v_{t-1}(l) A_{s_l,s_k} B_{s_k,x_t}$
- termination
  - $v_{\text{term}} = \max_k v_T(k)$
  - $c_{\text{term}} = \arg\max_k v_T(k)$
- c_{term} からスタートしてさかのぼれば、隠れ状態の列が得られる

Contents

観測データからパラメータの値を得る（推定）

- 観測データの尤度を計算するにしても…
- 観測データ列に対して最良の隠れ状態列を得るにしても…
- パラメータ A, B がすでに分かっている必要がある
- そこで…
- 混合分布の場合と同様、EM アルゴリズムにより、A と B の値を推定することにする

隠れマルコフモデルのEMアルゴリズム

- EMアルゴリズムの導出には、やはりJensenの不等式を使う

$$\ln p(\boldsymbol{x}_{1:T}; A, B) = \ln \sum_{\boldsymbol{z}_{1:T}} p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B)$$

$$= \ln \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \frac{p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B)}{q(\boldsymbol{z}_{1:T})} \geq \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \ln \frac{p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B)}{q(\boldsymbol{z}_{1:T})} \tag{33}$$

- EMアルゴリズムでは、この下界を最大化する

$$A, B = \arg\max_{A,B} \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \ln \frac{p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B)}{q(\boldsymbol{z}_{1:T})}$$

$$= \arg\max_{A,B} \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \ln p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B) \tag{34}$$

cf. http://cs229.stanford.edu/section/cs229-hmm.pdf

$$
\begin{aligned}
&\sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \ln p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; A, B) \\
&= \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \ln \left[\left(\prod_{t=1}^{T} p(x_t | z_t; B) \right) \left(\prod_{t=1}^{T} p(z_t | z_{t-1}; A) \right) \right] \\
&= \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \ln \left[\left(\prod_{t=1}^{T} B_{z_t, x_t} \right) \left(\prod_{t=1}^{T} A_{z_{t-1}, z_t} \right) \right] \\
&= \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \left[\sum_{t=1}^{T} \left(\ln B_{z_t, x_t} + \ln A_{z_{t-1}, z_t} \right) \right] \qquad (35)
\end{aligned}
$$

HMM の EM アルゴリズムの E step

- E step では、モデルパラメータを固定したうえで、観測データが所与のときの潜在変数の条件付き確率分布を求めるのだった（cf. 混合分布の講義）
- ということは、いま考えている HMM の場合、A と B を固定したうえで $p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}; A_{\text{old}}, B_{\text{old}})$ を計算して、それを $q(\boldsymbol{z}_{1:T})$ の解とするのが、E step となる
- 以下、 M step の説明をするが、そこでは $q(\boldsymbol{z}_{1:T})$ を $p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}; A_{\text{old}}, B_{\text{old}}) = \frac{p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A_{\text{old}}, B_{\text{old}})}{p(\boldsymbol{x}_{1:T}; A_{\text{old}}, B_{\text{old}})}$ で置き換えて計算を進める

HMM の EM アルゴリズムの M step (1/2)

最大化する関数は

$$
\begin{aligned}
\mathcal{L}(A, B, \{\lambda_k\}, \{\mu_k\}) &= \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \left(\ln B_{z_t, x_t} + \ln A_{z_{t-1}, z_t} \right) \\
&\quad + \sum_{l=1}^{K} \lambda_l \left(1 - \sum_{k=1}^{K} A_{s_l, s_k} \right) + \sum_{k=1}^{K} \mu_k \left(1 - \sum_{w=1}^{W} B_{s_k, v_w} \right) \\
&= \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{l=1}^{K} \sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \land z_t = s_k) \ln A_{s_l, s_k} \\
&\quad + \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{k=1}^{K} \sum_{w=1}^{W} \sum_{t=1}^{T} \mathbb{1}(z_t = s_k \land x_t = v_w) \ln B_{z_t, x_t} \\
&\quad + \sum_{l=1}^{K} \lambda_l \left(1 - \sum_{k=1}^{K} A_{s_l, s_k} \right) + \sum_{k=1}^{K} \mu_k \left(1 - \sum_{w=1}^{W} B_{s_k, v_w} \right) \qquad (36)
\end{aligned}
$$

$$\frac{\partial \mathcal{L}(A, B, \{\lambda_k\}, \{\mu_k\})}{\partial A_{s_l, s_k}} = \frac{\sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)}{A_{s_l, s_k}} - \lambda_l \tag{37}$$

$$\therefore A_{s_l, s_k} = \frac{\sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k)}{\sum_{k'=1}^{K} \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_{k'})} \tag{38}$$

$$\begin{aligned} \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k) &= \sum_{t=1}^{T} \sum_{\boldsymbol{z}_{1:T}} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k) q(\boldsymbol{z}_{1:T}) \\ &= \frac{1}{p(\boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}})} \sum_{t=1}^{T} \sum_{\boldsymbol{z}_{1:T}} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k) p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}}) \\ &= \frac{1}{p(\boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}})} \sum_{t=1}^{T} p(z_{t-1} = s_l, z_t = s_k, \boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}}) \end{aligned} \tag{39}$$

$p(z_{t-1} = s_l, z_t = s_k, \boldsymbol{x}_{1:T}; A, B)$ は、$\alpha_{t-1}(l) \equiv p(x_1, \ldots, x_{t-1}, z_{t-1} = s_l; A, B)$ と A_{s_l, s_k} と B_{s_k, x_t} と $p(x_{t+1}, \ldots, x_T | z_t = s_k; A, B)$ とが分かれば計算できる（次のスライド）。
$\beta_t(k) \equiv p(x_{t+1}, \ldots, x_T | z_t = s_k; A, B)$ とおく（これの求め方は後で）。

下の式変形で、一つ目の等号は、条件付き確率の定義を使っただけ。
二つ目の等号は、隠れマルコフモデルにおいて仮定していることに基づく。

$$
\begin{aligned}
&p(z_{t-1} = s_l, z_t = s_k, \boldsymbol{x}_{1:T}; A, B) \\
&= p(x_1, \ldots, x_{t-1}, z_{t-1} = s_l; A, B) \\
&\quad \times p(z_t = s_k | x_1, \ldots, x_{t-1}, z_{t-1} = s_l; A, B) \\
&\quad \times p(x_t | x_1, \ldots, x_{t-1}, z_{t-1} = s_l, z_t = s_k; A, B) \\
&\quad \times p(x_{t+1}, \ldots, x_T | x_1, \ldots, x_{t-1}, x_t, z_{t-1} = s_l, z_t = s_k; A, B) \\
&= p(x_1, \ldots, x_{t-1}, z_{t-1} = s_l; A, B) \\
&\quad \times p(z_t = s_k | z_{t-1} = s_l; A) p(x_t | z_t = s_k; B) \\
&\quad \times p(x_{t+1}, \ldots, x_T | z_t = s_k; A, B) \\
&= \alpha_{t-1}(l) A_{s_l, s_k} B_{s_k, x_t} \beta_t(k)
\end{aligned} \tag{40}
$$

ただし、$t = 1$ の場合は別扱い。$l \neq 0$ のときは $p(z_0 = s_l, z_t = s_k, \boldsymbol{x}_{1:T}; A, B) = 0$ で、$l = 0$ のときは $p(z_0 = s_l, z_t = s_k, \boldsymbol{x}_{1:T}; A, B) = A_{s_0, s_k} B_{s_k, x_t} \beta_t(k)$ となる。
$\beta_t(k)$ の求め方は次のスライド以降で説明する。

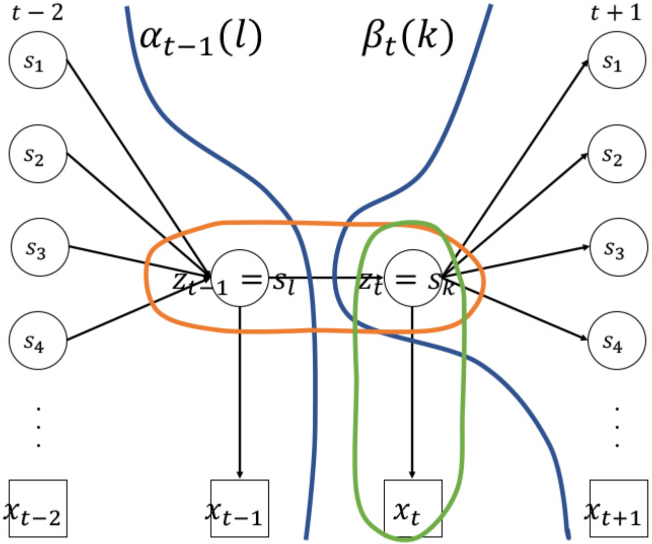


Figure: パラメータ A を求める計算の概念図

Backward procedure

$\beta_t(k)$ は、動的計画法の一種である backward procedure によって、効率的に計算できる

cf. https://web.stanford.edu/~jurafsky/slp3/A.pdf

$$\beta_{T-1}(k) = p(x_T|z_{T-1} = s_k) = \sum_{z_T \in \mathcal{S}} p(z_T|z_{T-1} = s_k)p(x_T|z_T) = \sum_{l=1}^{K} A_{s_k,s_l} B_{s_l,x_T} \tag{41}$$

$$\begin{aligned}\beta_{T-2}(k) &= p(x_{T-1}, x_T|z_{T-2} = s_k) \\ &= \sum_{z_{T-1} \in \mathcal{S}} p(z_{T-1}|z_{T-2} = s_k)p(x_{T-1}|z_{T-1})p(x_T|z_{T-1}) \\ &= \sum_{l=1}^{K} A_{s_k,s_l} B_{s_l,x_{T-1}} \beta_{T-1}(l) \end{aligned} \tag{42}$$

同様に考えて

$$\beta_t(k) = \sum_{l=1}^{K} A_{s_k,s_l} B_{s_l,x_{t+1}} \beta_{t+1}(l) \tag{43}$$

Forward-backward algorithm

- 隠れマルコフモデルについて、$\alpha_t(k)$ と $\beta_t(k)$ を上述の方法で求めるアルゴリズムを forward-backward アルゴリズムと呼ぶ

1. $\alpha_t(k)$ を求めるのが forward procedure
   - A, B が既知なら、forward procedure だけでデータ尤度 $p(\boldsymbol{x}; A, B)$ を求めることができた（cf. 式 (12)）
2. $\beta_t(k)$ を求めるのが backward procedure

HMMのEMアルゴリズムのM step (2/2)

$\sum_{\boldsymbol{z}_{1:T}} \mathbb{1}(z_{t-1} = s_l \wedge z_t = s_k) p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}}) = \alpha_{t-1}(l) A_{s_l,s_k} B_{s_k,x_t} \beta_t(k)$ より

$$A_{s_l,s_k} \propto \sum_{t=1}^{T} \alpha_{t-1}(l) A_{s_l,s_k} B_{s_k,x_t} \beta_t(k) \tag{44}$$

$$\therefore A_{s_l,s_k} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(l) A_{s_l,s_k} B_{s_k,x_t} \beta_t(k)}{\sum_{k'=1}^{K} \sum_{t=1}^{T} \alpha_{t-1}(l) A_{s_l,s_{k'}} B_{s_{k'},x_t} \beta_t(k')} \tag{45}$$

次は B_{s_k,v_w} を求める。

$$\frac{\partial \mathcal{L}(A, B, \{\lambda_k\}, \{\mu_k\})}{\partial B_{s_k,v_w}} = \frac{\sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_t = s_k \wedge x_t = v_w)}{B_{z_t,x_t}} - \mu_k \tag{46}$$

$$B_{s_k,v_w} = \frac{\sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_t = s_k \wedge x_t = v_w)}{\sum_{w'=1}^{W} \sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_t = s_k \wedge x_t = v_{w'})} \tag{47}$$

$$
\begin{aligned}
\sum_{\boldsymbol{z}_{1:T}} q(\boldsymbol{z}_{1:T}) \sum_{t=1}^{T} \mathbb{1}(z_t = s_k \wedge x_t = v_w) &= \sum_{t=1}^{T} \sum_{\boldsymbol{z}_{1:T}} \mathbb{1}(z_t = s_k \wedge x_t = v_w) q(\boldsymbol{z}_{1:T}) \\
&= \frac{1}{p(\boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}})} \sum_{t=1}^{T} \sum_{\boldsymbol{z}_{1:T}} \mathbb{1}(z_t = s_k \wedge x_t = v_w) p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}}) \\
&= \frac{1}{p(\boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}})} \sum_{t=1}^{T} \mathbb{1}(x_t = v_w) p(z_t = s_k, \boldsymbol{x}_{1:T}; A_{\mathsf{old}}, B_{\mathsf{old}}) \qquad (48)
\end{aligned}
$$

$$
\begin{aligned}
p(z_t = s_k, \boldsymbol{x}_{1:T}; A, B) &= p(x_1, \ldots, x_t, z_t = s_k; A, B) p(x_{t+1}, \ldots, x_T | x_1, \ldots, x_t, z_t = s_k; A, B) \\
&= p(x_1, \ldots, x_t, z_t = s_k; A, B) p(x_{t+1}, \ldots, x_T | z_t = s_k; A, B) \\
&= \alpha_t(k) \beta_t(k) \qquad (49)
\end{aligned}
$$

$$
\therefore B_{s_k, v_w} = \frac{\sum_{t=1}^{T} \mathbb{1}(x_t = v_w) \alpha_t(k) \beta_t(k)}{\sum_{w'=1}^{W} \sum_{t=1}^{T} \mathbb{1}(x_t = v_{w'}) \alpha_t(k) \beta_t(k)} = \frac{\sum_{t=1}^{T} \mathbb{1}(x_t = v_w) \alpha_t(k) \beta_t(k)}{\sum_{t=1}^{T} \alpha_t(k) \beta_t(k)} \qquad (50)
$$

Contents

条件付き独立性

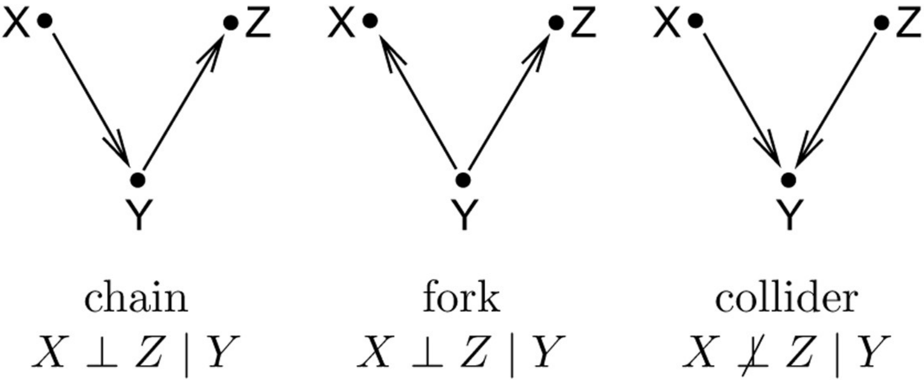


Figure: 条件付き独立性

cf. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-S6-S5
chain は head-to-tail 型、fork は tail-to-tail 型、collider は head-to-head 型とも呼ぶ。

条件付き独立性の証明

- chain について $p(X,Z|Y)=p(X|Y)p(Z|Y)$ を下に示す
  - chain の図が表すのは $p(X,Y,Z)=p(Z|Y)p(Y|X)p(X)$ という分解
  - よって、$p(X,Z|Y)=\frac{p(X,Y,Z)}{p(Y)}=\frac{p(Z|Y)p(Y|X)p(X)}{p(Y)}=p(Z|Y)p(X|Y)$
- fork について $p(X,Z|Y)=p(X|Y)p(Z|Y)$ を下に示す
  - fork の図が表すのは $p(X,Y,Z)=p(X|Y)p(Z|Y)p(Y)$ という分解
  - よって、$p(X,Z|Y)=\frac{p(X,Y,Z)}{p(Y)}=\frac{p(X|Y)p(Z|Y)p(Y)}{p(Y)}=p(X|Y)p(Z|Y)$
- collider について $p(X,Z|Y)\neq p(X|Y)p(Z|Y)$ を下に示す
  - collider の図が表すのは $p(X,Y,Z)=p(X)p(Z)p(Y|X,Z)$ という分解
  - このとき、$p(X,Z|Y)=\frac{p(X,Y,Z)}{p(Y)}=\frac{p(X)p(Z)p(Y|X,Z)}{p(Y)}\neq p(X|Y)p(Z|Y)$

5つ目の課題（1月31日23:59まで）

- コインがある。このコインの表が出る確率を μ とする。
- μ の事前分布は、下の密度関数を持つベータ分布だとする。

$$f(\mu;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{51}$$

問1 $\int_0^1 \mu f(\mu;\alpha,\beta)d\mu$ を計算せよ。

問2 このコインを投げたところ、表が m 回出て、裏が l 回出た。このとき、μ の事後分布を求めよ。