HMM の学習

1　HMM の学習

HMM の学習とは，観測系列 $o_1^T = o_1, o_2, \ldots, o_T$ の対数尤度 $\log P(o_1^T|\lambda)$ を最大にするような HMM のパラメータ $\lambda = \{A, B\}$ を求めることである．ここで，A は状態遷移確率（$A = a_{11}, \cdots, a_{ij}, \cdots, a_{NN}$），$B$ は生成確率（$B = b_i(o)$）を表している．また，状態には 1〜N を想定する．

HMM では，観測系列 o_1^T を生成した状態系列 $q_1^T = q_1, q_2, \ldots, q_T$ が非観測である．つまり，各観測データ o_t を生成した状態 q_t が 1〜N のうちどれか分からない．また，観測系列に対して，内部状態がどう遷移していったかも見えない．そのため，どの状態がどの状態に何回遷移したかや，どの状態が何回観測データ o を生成したか等の頻度がカウントできず，単純に相対頻度を使って a_{ij} と $b_i(o)$ を決めることができない．

そこで HMM では，EM アルゴリズムを使ってパラメータ λ を推定する．EM アルゴリズムでは，現在与えられているパラメータ λ よりも観測系列の対数尤度が大きくなるパラメータ $\bar{\lambda}$ を探すことを繰り返し，観測系列の対数尤度が最大となるパラメータを見つける．

2　Q 関数

観測データ o_t について，パラメータ λ を $\bar{\lambda}$ に更新したときの対数尤度の差は以下のようになる（詳細は Appendix）．

$$\log P(o_t|\bar{\lambda}) - \log P(o_t|\lambda) = \sum_{q_t} P(q_t|o_t,\lambda)\log\frac{P(o_t,q_t|\bar{\lambda})}{P(o_t,q_t|\lambda)} + \sum_{q_t} P(q_t|o_t,\lambda)\log\frac{P(q_t|o_t,\lambda)}{P(q_t|o_t,\bar{\lambda})} \tag{1}$$

ここで，右辺第 2 項の

$$\sum_{q_t} P(q_t|o_t,\lambda)\log\frac{P(q_t|o_t,\lambda)}{P(q_t|o_t,\bar{\lambda})}$$

は KL ダイバージェンス

$$D_{KL}(P||Q) = \sum_{x} P(X=x)\log\frac{P(X=x)}{Q(X=x)}$$

の形になっており，非負性が保証されている．

$$\sum_{q_t} P(q_t|o_t,\lambda)\log\frac{P(q_t|o_t,\lambda)}{P(q_t|o_t,\bar{\lambda})} \geq 0 \tag{2}$$

そのため，

$$\begin{aligned}\log P(o_t|\bar{\lambda}) - \log P(o_t|\lambda) &\geq \sum_{q_t} P(q_t|o_t,\lambda)\log\frac{P(o_t,q_t|\bar{\lambda})}{P(o_t,q_t|\lambda)} \\ &= \sum_{q_t} P(q_t|o_t,\lambda)\log P(o_t,q_t|\bar{\lambda}) - \sum_{q_t} P(q_t|o_t,\lambda)\log P(o_t,q_t|\lambda)\end{aligned}$$

ここで，

$$Q(\bar{\lambda},\lambda) = \sum_{q_t} P(q_t|o_t,\lambda)\log P(o_t,q_t|\bar{\lambda}) \tag{3}$$

とおくと，

$$\log P(o_t|\bar{\lambda}) - \log P(o_t|\lambda) \geq Q(\bar{\lambda},\lambda) - Q(\lambda,\lambda) \tag{4}$$

となる．

(4) 式は，$Q(\bar{\lambda},\lambda) > Q(\lambda,\lambda)$ となるような $\bar{\lambda}$ を見つければ，自動的に $\log P(o_t|\bar{\lambda}) - \log P(o_t|\lambda) > 0$ となり，観測データ o_t に対する対数尤度を増加させることができることを表している．EM アルゴリズムでは，適当なパラメータ λ から開始し，値が収束するまで $Q(\bar{\lambda},\lambda)$ を最大化する $\bar{\lambda}$ を求めることを繰り返す．この関数 Q を Q 関数といい，観測系列 $o_1^T = o_1$，o_2，... ，o_T について拡張すると，

$$Q(\bar{\lambda},\lambda) = \frac{1}{P(o_1^T|\lambda)}\sum_{q_1^T} P(o_1^T,q_1^T|\lambda)\log P(o_1^T,q_1^T|\bar{\lambda}) \tag{5}$$

となる（詳細は Appendix）．

「現状のパラメータよりも対数尤度を高くできるパラメータの内で，最も対数尤度を大きくするものを繰り返し選んでいけば，対数尤度を最大とするパラメータが見つかるだろう」というのが EM アルゴリズムの戦略である．ただし，どのような値に収束するかは初期値の与え方に大きく依存し，収束する値は一般に局所的最大値にすぎない．

3 Q 関数の最大化

式 (5) を最大化するパラメータを求めるために，まずは式 (5) を変形する．具体的には，a_{ij} と $b_i(o)$ についてそれぞれ微分しやすいような形に変形する（導出の詳細は Appendix を参照)．

$$\begin{aligned}Q(\bar{\lambda},\lambda) &= \frac{1}{P(o_1^T|\lambda)}\sum_{q_1^T} P(o_1^T,q_1^T|\lambda)\log P(o_1^T,q_1^T|\bar{\lambda}) \\ &= \sum_i\sum_j d_{ij}\log\bar{a}_{ij} + \sum_j\sum_k e_{jk}\log\bar{b}_j(k)\end{aligned}$$

ここで，Q 関数の各項は独立である（各項に各パラメータただ 1 つがある）ため，各項をそれぞれ $\sum_j a_{ij} = 1$，$\sum_k b_j(k) = 1$ の下で最大化すればいい．

ここで，最大化すべき関数はいずれも下記の形をしているが，

$$f(x) = \sum_i a_i \log\ x_i \qquad (a_i > 0 \text{ かつ } \sum_i x_i = 1)$$

この関数を最大化する x_i はラグランジュの未定乗数法より，次式で与えられる (詳細は「確率的言語モデル」の p.118 を参照)．

$$x_i = \frac{a_i}{\sum_i a_i}$$

したがって，Q 関数を最大化するパラメータを以下のように得る．

$$\begin{aligned}\bar{a}_{ij} &= \frac{d_{ij}}{\sum_j d_{ij}} \\ &= \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_j \sum_{t=1}^{T} \xi_t(i,j)} \\ &= \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \sum_j \xi_t(i,j)} \end{aligned} \tag{6}$$

$$\begin{aligned}\bar{b}_j(k) &= \frac{e_{jk}}{\sum_k e_{jk}} \\ &= \frac{\sum_{t|o_t=k} \gamma_t(j)}{\sum_k \sum_{t|o_t=k} \gamma_t(j)} \\ &= \frac{\sum_{t|o_t=k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \end{aligned} \tag{7}$$

forward backward アルゴリズムは，式 (6)(7) における ξ と γ を効率的に計算するのに使用される．

Appendix

式 (1) の導出

$$\begin{aligned}\log P(o_t|\bar{\lambda}) - \log P(o_t|\lambda) &= \log \frac{P(o_t|\bar{\lambda})}{P(o_t|\lambda)} \\ &= \log \frac{P(o_t|\bar{\lambda})}{P(o_t|\lambda)} \ \times \ 1 \\ &= \log \frac{P(o_t|\bar{\lambda})}{P(o_t|\lambda)} \sum_{q_t} P(q_t|o_t,\lambda) \\ &= \sum_{q_t} P(q_t|o_t,\lambda) \log \frac{P(o_t|\bar{\lambda})}{P(o_t|\lambda)}\end{aligned}$$

$P(X|Y,Z) = \frac{P(X,Y|Z)}{P(Y|Z)}$ より，$P(Y|Z) = \frac{P(X,Y|Z)}{P(X|Y,Z)}$ だから，

$$\begin{aligned}\log P(o_t|\bar{\lambda}) - \log P(o_t|\lambda) &= \sum_{q_t} P(q_t|o_t,\lambda) \log \left\{ \frac{P(o_t,q_t|\bar{\lambda})}{P(o_t,q_t|\lambda)} \frac{P(q_t|o_t,\lambda)}{P(q_t|o_t,\bar{\lambda})} \right\} \\ &= \sum_{q_t} P(q_t|o_t,\lambda) \log \frac{P(o_t,q_t|\bar{\lambda})}{P(o_t,q_t|\lambda)} + \sum_{q_t} P(q_t|o_t,\lambda) \log \frac{P(q_t|o_t,\lambda)}{P(q_t|o_t,\bar{\lambda})}\end{aligned}$$

式（５）の導出

$$\begin{aligned} Q(\bar{\lambda}, \lambda) &= \sum_{t=1}^{T} \sum_{q_t} P(q_t|o_t, \lambda) \log P(o_t, q_t|\bar{\lambda}) \\ &= \sum_{q_1^T} P(q_1^T|o_1^T, \lambda) \log P(o_1^T, q_1^T|\bar{\lambda}) \end{aligned}$$

$P(X|Y,Z) = \frac{P(X,Y|Z)}{P(Y|Z)}$ より，

$$\begin{aligned} Q(\bar{\lambda}, \lambda) &= \sum_{q_1^T} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \log P(o_1^T, q_1^T|\bar{\lambda}) \\ &= \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \log P(o_1^T, q_1^T|\bar{\lambda}) \end{aligned} \tag{8}$$

式（６）の導出

$$Q(\bar{\lambda}, \lambda) = \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \log P(o_1^T, q_1^T|\bar{\lambda})$$

$P(X|Y,Z) = \frac{P(X,Y|Z)}{P(Y|Z)}$ より，$P(X,Y|Z) = P(Y|Z)P(X|Y,Z)$ だから，

$$\begin{aligned} Q(\bar{\lambda}, \lambda) &= \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \log \Big\{ P(q_1^T|\bar{\lambda}) P(o_1^T|q_1^T, \bar{\lambda}) \Big\} \\ &= \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \Big\{ \log P(q_1^T|\bar{\lambda}) + \log P(o_1^T|q_1^T, \bar{\lambda}) \Big\} \end{aligned} \tag{9}$$

ここでまず，右辺の $P(q_1^T|\bar{\lambda})$ は，パラメータ $\bar{\lambda}$ が既知の状態での q_1^T の確率であるから，すなわち，

$$P(q_1^T|\bar{\lambda}) = \prod_{t=1}^{T-1} \bar{a}_{q_t q_{t+1}}$$

ここで，１次のマルコフ性が仮定されていることに注意してほしい．また簡単のため，初期状態 0（t=0）からの遷移や，終了状態 F(t=T+1) への遷移はここでは考慮していない．

つぎに，同じく右辺の $P(o_1^T|q_1^T, \bar{\lambda})$ は，パラメータ $\bar{\lambda}$ と内部の状態遷移 q_1^T が既知の状態での o_1^T の確率であるから，すなわち，

$$P(o_1^T|q_1^T, \bar{\lambda}) = \prod_{t=1}^{T} \bar{b}_{q_t}(o_t)$$

以上を式 (9) へ代入すると，

$$
\begin{aligned}
Q(\bar{\lambda}, \lambda) &= \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \left\{ \log \prod_{t=1}^{T-1} \bar{a}_{q_t q_{t+1}} + \log \prod_{t=1}^{T} \bar{b}_{q_t}(o_t) \right\} \\
&= \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \left\{ \sum_{t=1}^{T-1} \log\ \bar{a}_{q_t q_{t+1}} + \sum_{t=1}^{T} \log\ \bar{b}_{q_t}(o_t) \right\} \\
&= \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \sum_{t=1}^{T-1} \log\ \bar{a}_{q_t q_{t+1}} + \frac{1}{P(o_1^T|\lambda)} \sum_{q_1^T} P(o_1^T, q_1^T|\lambda) \sum_{t=1}^{T} \log\ \bar{b}_{q_t}(o_t) \\
&= \sum_{q_1^T} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \sum_{t=1}^{T-1} \log\ \bar{a}_{q_t q_{t+1}} + \sum_{q_1^T} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \sum_{t=1}^{T} \log\ \bar{b}_{q_t}(o_t) \\
&= \sum_{q_1^T} \sum_{t=1}^{T-1} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{a}_{q_t q_{t+1}} + \sum_{q_1^T} \sum_{t=1}^{T} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{b}_{q_t}(o_t)
\end{aligned}
\tag{10}
$$

式 (10) の第 1 項は，あらゆる q_1^T について足しこまずとも「$q_t = i$，$q_{t+1} = j$」ごとにまとめることができる（つまり，$q_t = i$，$q_{t+1} = j$ であるような q_1^T について，$P(o_1^T, q_1^T|\lambda)$ を全て足しこんでしまっておく）．

$$
\sum_{q_1^T} \sum_{t=1}^{T-1} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{a}_{q_t q_{t+1}} = \sum_{i} \sum_{j} \sum_{t=1}^{T-1} \frac{P(o_1^T, q_t = i, q_{t+1} = j|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{a}_{ij} \tag{11}
$$

ここで，

$$
\xi_t(i, j) = P(q_t = i, q_{t+1} = j|o_1^T, \lambda)
$$

とおくと，$P(X|Y, Z) = \frac{P(X,Y|Z)}{P(Y|Z)}$ より，

$$
\xi_t(i, j) = \frac{P(o_1^T, q_t = i, q_{t+1} = j|\lambda)}{P(o_1^T|\lambda)}
$$

であり，さらに，

$$
d_{ij} = \sum_{t=1}^{T-1} \xi_t(i, j) \tag{12}
$$

とおくと，式 (11) は，

$$
\begin{aligned}
\sum_{q_1^T} \sum_{t=1}^{T-1} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{a}_{q_t q_{t+1}} &= \sum_{i} \sum_{j} \sum_{t=1}^{T-1} \xi_t(i, j) \log\ \bar{a}_{ij} \\
&= \sum_{i} \sum_{j} d_{ij} \log\ \bar{a}_{ij}
\end{aligned}
\tag{13}
$$

となる．

同様に，式 (10) の第 2 項も，あらゆる q_1^T について足しこまずとも「$q_t = j$」ごとにまとめることができる（つまり，$q_t = j$ であるような q_1^T について，$P(o_1^T, q_1^T|\lambda)$ を全て足しこんでしまっておく）．

$$
\sum_{q_1^T} \sum_{t=1}^{T} \frac{P(o_1^T, q_1^T|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{b}_{q_t}(o_t) = \sum_{j} \sum_{t=1}^{T} \frac{P(q_t = j, o_1^T|\lambda)}{P(o_1^T|\lambda)} \log\ \bar{b}_j(o_t) \tag{14}
$$

また，$t=1\sim T$ についての足しこみを $o_t=k$ となるような t の，k についての足しこみに変更すると，

$$\sum_{q_1^T}\sum_{t=1}^{T}\frac{P(o_1^T,q_1^T|\lambda)}{P(o_1^T|\lambda)}\log\ \bar{b}_{q_t}(o_t)=\sum_j\sum_k\sum_{t|o_t=k}\frac{P(q_t=j,o_1^T|\lambda)}{P(o_1^T|\lambda)}\log\ \bar{b}_j(k)$$

ここで，

$$\gamma_t(j)=P(q_t=j|o_1^T,\lambda)$$

とおくと，$P(X|Y,Z)=\frac{P(X,Y|Z)}{P(Y|Z)}$ より，

$$\gamma_t(j)=\frac{P(q_t=j,o_1^T|\lambda)}{P(o_1^T|\lambda)}$$

であり，さらに

$$e_{jk}=\sum_{t|o_t=k}\gamma_t(j) \tag{15}$$

とおくと，式 (14) は，

$$\begin{aligned}\sum_{q_1^T}\sum_{t=1}^{T}\frac{P(o_1^T,q_1^T|\lambda)}{P(o_1^T|\lambda)}\log\ \bar{b}_{q_t}(o_t)&=\sum_j\sum_k\sum_{t|o_t=k}\gamma_t(j)\ \log\ \bar{b}_j(k)\\&=\sum_j\sum_k e_{jk}\ \log\ \bar{b}_j(k)\end{aligned} \tag{16}$$

となる．

式 (13) と式 (16) より，式 (10) は，

$$Q(\bar{\lambda},\lambda)=\sum_i\sum_j d_{ij}\log\ \bar{a}_{ij}+\sum_j\sum_k e_{jk}\ \log\ \bar{b}_j(k)$$

となる．