東京大学 工学部 計数工学科/物理工学科

応用音響学：Baum-Welch アルゴリズム

嵯峨山 茂樹 ＜sagayama@hil.t.u-tokyo.ac.jp＞
東京大学 工学部 計数工学科 http://hil.t.u-tokyo.ac.jp/

EMアルゴリズムとBaum-Welchアルゴリズム

- 内容
  - HMMの3つの問題
  - EMアルゴリズム
  - EMアルゴリズム収束性の証明
  - Baum-Weichアルゴリズム
- 参考文献
  - 中川聖一「確率モデルによる音声認識」電子情報通信学会, 1988.
  - Lawrence Rabiner他「音声認識の基礎（下）」NTTアドバンステクノロジ株式会社, 1995.
  - 上坂吉則他「パターン認識と学習のアルゴリズム」文一総合出版, 1990.
  - 北研二他「音声言語処理 コーパスに基づくアプローチ」森北出版株式会社, 1996.

HMMの3つの基本問題

モデル $\lambda = (\overset{\text{状態遷移確率行列}}{A}, \overset{\text{観測シンボル確率分布}}{B}, \overset{\text{初期状態分布}}{\pi})$、
観測系列 $\mathrm{Y} = (\mathrm{y}_1 \mathrm{y}_2 \cdots \mathrm{y}_T)$

問題1 モデル λ に対する観測系列 Y の確率 $P(\mathrm{Y}|\lambda)$ の計算

- モデル λ が観測系列 Y に対してどの程度適応しているか
  - … ex. Forwardアルゴリズム

問題2 最適な状態系列 $\mathrm{q} = (q_1 q_2 \cdots q_T)$ の発見

- 観測系列 Y がどの状態系列 q から生成されたと考えられるか
  - … ex. Viterbiアルゴリズム

問題3 $P(\mathrm{Y}|\lambda)$ を最大とするようなモデルパラメータ λ の調整

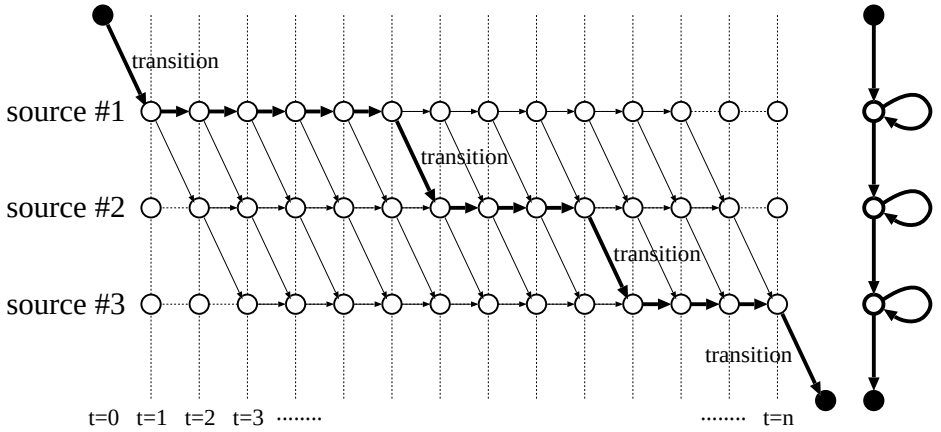
- 観測系列 Y を生成するためのパラメータ λ の最適化
  - … ex. EMアルゴリズム、Baum-Welchアルゴリズム

Viterbi 学習と Baum-Welch 学習

■ パラメータ学習アルゴリズム

- **Viterbi 学習アルゴリズム:**
  確率最大の経路に沿って再学習の繰り返し

- **Baum-Welch 学習アルゴリズム:**
  全経路を考えてトレリス各点で確率重みつき再学習の繰り返し

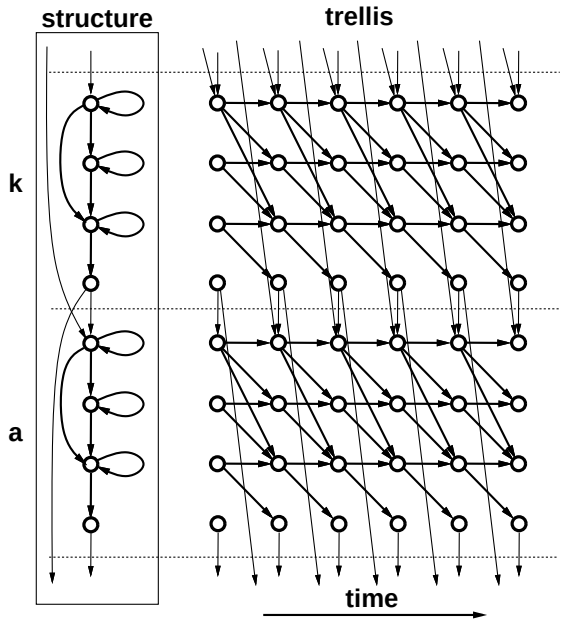


HMMの学習

学習アルゴリズム

- Baum-Welch アルゴリズム(forward-backward algorithm)
- Viterbi 学習アルゴリズム

HMM トレリス計算(縦方向は状態、横方向は時間)



Trellis Computation

図1. HMM トレリス計算(縦方向は状態、横方向は時間)

EM(Expectation-Maximization)アルゴリズム

観測できるデータx:不完全データ
観測できないデータy } 完全データ

完全なデータが得られる場合は最尤推定によってパラメータを推定することができるが、不完全なデータしか得られない場合は最尤推定を直接行なうことはできない。
EMアルゴリズムは、不完全データからの対数尤度を最大化するために、完全データからの対数尤度の**期待値を最大化**する。

EMアルゴリズム

y: 不完全な観測データ、x: yに付随して得られる付加データ、ϕ: 初期パラメータ、$f(y|\theta)$, $f(x|\theta)$: 確率密度関数 とする。

尤度関数$P(\hat{\theta}; y)$にxを導入

$$P(\hat{\theta}; y) = \frac{P(\hat{\theta}; x, y)}{P(\hat{\theta}; x, y)P(\hat{\theta}; y)} = \frac{P(\hat{\theta}; x, y)}{P(\hat{\theta}; x|y)} \tag{1}$$

対数尤度

$$\log P(\hat{\theta}; y) = \log P(\hat{\theta}; x, y) - \log P(\hat{\theta}; x|y) \tag{2}$$

EMアルゴリズム

確率変数Xのすべてに渡って、あらかじめ設定されている推定値θを使って期待値をとる。

$$
\begin{aligned}
E[\log P(\hat{\theta};y)|x] &= E[\textstyle\sum_x P(\theta,x|y)\log P(\hat{\theta};y)] && (3)\\
&= \log P(\hat{\theta};y) && (4)\\
E[\log P(\hat{\theta};y)] &= E[\log P(\hat{\theta};x,y) - \log P(\hat{\theta};x|y)] && (5)\\
&= E[\log P(\hat{\theta};x,y)] - E[\log P(\hat{\theta},x|y)] && (6)
\end{aligned}
$$

これから

$$\log P(\hat{\theta};y) = E[\log P(\hat{\theta};x,y)] - E[\log P(\hat{\theta};x|y)] \quad (7)$$

が得られる。

EMアルゴリズム

ここで、

$$\log P(\hat{\theta};y) = E[\log P(\hat{\theta};x,y)] - E[\log P(\hat{\theta};x|y]) \quad (8)$$

$$= \sum_x P(\theta;x,y)\log P(\hat{\theta};x,y) - \sum_x P(\theta;x|y)\log P(\hat{\theta};x|y) \quad (9)$$

$$= Q(\theta,\hat{\theta}) - H(\theta,\hat{\theta}) \quad (10)$$

Jensenの不等式より、
$H(\theta,\hat{\theta}) \leq H(\theta,\hat{\theta})$なので、$Q(\theta,\hat{\theta}) \geq Q(\theta,\theta)$となるように$\hat{\theta}$を設定すれば$P(y|\hat{\theta}) \geq P(y|\hat{\theta})$となる。
つまり、Qを最大化するようθを変えるとPも最大化される。
EMアルゴリズムは$P(y|\theta)$を直接最大化する代わりにQを最大化していくアルゴリズムである。

EMアルゴリズム

Jensenの不等式

$$H(\Phi,\Phi) \geq H(\Phi,\overline{\Phi})$$

$$\begin{aligned} H(\Phi,\Phi) &= \int \log f(y|x,\Phi) f(y|x,\Phi) dy \\ H(\Phi,\overline{\Phi}) &= \int \log f(y|x,\overline{\Phi}) f(y|x,\Phi) dy \end{aligned}$$

$$\begin{aligned} H(\Phi,\overline{\Phi}) - H(\Phi,\Phi) &= \int (\log f(y|x,\overline{\Phi}) - \log f(y|x,\Phi)) f(y|x,\Phi) dy \\ &= \int \log \frac{f(y|x,\overline{\Phi})}{f(y|x,\Phi)} f(y|x,\Phi) dy \\ &\leq \int (\frac{f(y|x,\overline{\Phi})}{f(y|x,\Phi)} - 1) f(y|x,\Phi) dy \quad \text{since} \quad \log A \leq A - 1 \\ &= \int f(y|x,\overline{\Phi}) dy - \int f(y|x,\Phi) dy = 0 \end{aligned}$$

等号は$f(y|x,\overline{\Phi}) = f(y|x,\Phi)$の時成立

EMアルゴリズム

これを観測可能な系列$X = \{x_1, \cdots, x_n\}$と観測不可能な系列$Y = \{y_1, \cdots, y_n\}$の場合に拡張すると以下のようになる。

$$L(X, \theta) = \sum_{k=1}^{n} \log f(x_k|\theta) \tag{11}$$

$$E[L(X, \hat{\theta})|X, \theta] = E[\sum_{k=1}^{n} \log f(x_k|\theta)|X, \theta] \tag{12}$$

$$= L(X, \theta) \tag{13}$$

$$Q(\theta, \hat{\theta}) = \sum_{k=1}^{n} Q_k(\theta, \hat{\theta}) \tag{14}$$

$$H(\theta, \hat{\theta}) = \sum_{k=1}^{n} H_k(\theta, \hat{\theta}) \tag{15}$$

EMアルゴリズム

実行手順

1. パラメータΦの初期値を設定
2. $Q(\Phi, \overline{\Phi})$を求める。
3. $Q(\Phi, \overline{\Phi})$を最大にするような$\overline{\Phi}$を選ぶ。
4. $\overline{\Phi}$をΦに設定し、収束条件が満たされなければ 2. へ、満たされれば終了

ステップ2. は**期待値操作**(Expectation step)、ステップ3. は**最大値操作**(Maximization step)と呼ばれている。

Baum-Welchアルゴリズム

- 観測系列の確率を最大化するモデルパラメータを解析的に直接求める方法は知られていない
- しかし尤度が局所的に最大になるモデルパラメータを求めることはできる
- EMアルゴリズムをHMMのパラメータ推定に適用…**Baum-Welchアルゴリズム**(Forward-Backwardアルゴリズム)

パラメータの推定

モデルと観測系列が与えられたとき、時刻tのとき状態iに存在し、時刻$t+1$のとき状態jに存在する確率$\xi_t(i,j)$を定義する。

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

上式の条件を満足するパスを以下に示す。

Baum-Welchアルゴリズム

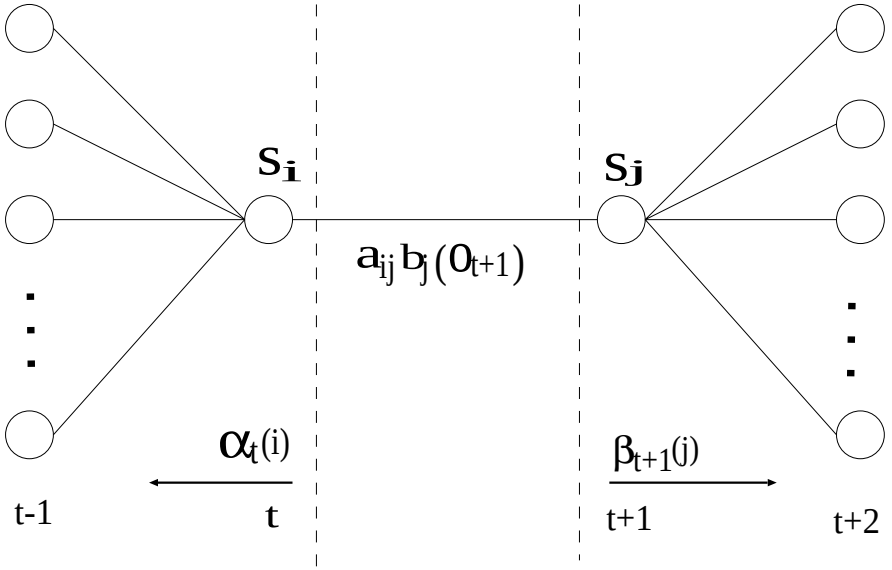


図2. 計算手順

前向き変数をα、後向き変数をβとすると$\xi_t(i,j)$は次のように書ける。

$$\xi_t(i,j) = \frac{P(q_t = i,\ q_{t+1} = j, Y|\lambda)}{P(Y|\lambda)}$$
$$= \frac{\alpha_t(i) a_{ij} b_j(y_{t+1}) \beta_{t+1}(j)}{P(Y|\lambda)}$$

Baum-Welchアルゴリズム

$$= \frac{\alpha_t(i)a_{ij}b_j(y_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(y_{t+1})\beta_{t+1}(j)}$$

$\gamma_t(i)$をモデルと観測系列全体が与えられたときに時刻tで状態iに存在する確率と定義すると、$\gamma_t(i)$は$\xi_t(i,j)$をjについて総和したものと考えられるので、

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j)$$

$\gamma_t(i)$をtについて総和をとると状態iを訪れた回数の期待値とみなせ、それから$t = T$を除いた値はiから遷移する回数の期待値とみなせる。

$\sum_{t=1}^{T-1}\gamma_t(i)$：$Y$において状態$i$から遷移する回数の期待値

同様に

$\sum_{t=1}^{T-1}\xi_t(i,j)$：$Y$において状態$i$から状態$j$へ遷移する回数の期待値

とみなせる。

Baum-Welchアルゴリズム

これらにより、HMMのパラメータ $\overset{\text{初期状態分布}}{\pi}$, A , $\overset{\text{観測シンボル確率分布}}{B}$ **の再推定式は以下のようになる。**

$$\overline{\pi_i} = \text{時刻}(t=1)\text{に状態}i\text{に存在すると期待される頻度(回数)} \tag{16}$$

$$= \gamma_1(i) \tag{17}$$

$$\overline{a_{ij}} = \frac{\text{状態}i\text{から状態}j\text{へ遷移する回数の期待値}}{\text{状態}i\text{から遷移する回数の期待値}} \tag{18}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{19}$$

$$\overline{b_j(k)} = \frac{\text{状態}j\text{にとどまりシンボル}v_k\text{を観測する回数の期待値}}{\text{状態}j\text{にとどまる回数の期待値}} \tag{20}$$

$$= \frac{\sum_{t=1 \ s.t. \ o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{21}$$

Baum-Welchアルゴリズム

現在のモデルを$\lambda = (A, B, \pi)$、(17),(18),(20) 式の左辺によって決定される再推定モデルを$\overline{\lambda} = (\overline{A}, \overline{B}, \pi)$とすると、

1. 初期モデルが尤度関数の臨界点($\overline{\lambda} = \lambda$ の点) を定義する、あるいは
2. $\overline{\lambda}$がλよりも$P(Y|\overline{\lambda}) > P(Y|\lambda)$の意味でより尤もらしい

つまり、観測系列が生成された可能性がより高い新しいモデル$\overline{\lambda}$を手に入れることができる。

このように$\overline{\lambda}$をλに入れ換えて繰り返し使いながら再推定計算を繰り返せばYがそのモデルから観測されたという確率をある点まで高めることができるが、これはあくまでも極大点である。

Baum-Welchアルゴリズム

(17),(19),(21)式の再推定式は次の補助関数を最大化することで直接求めることができる。

$$Q(\lambda', \lambda) = \sum_q P(Y, q|\lambda') \log P(Y, q|\lambda)$$

なぜなら

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow P(Y|\lambda) \geq P(Y|\lambda')$$

の関係があるので$P(Y|\lambda)$を増加させるという意味でλに関する関数$Q(\lambda', \lambda)$をλ'を改善しながら最大化できるからである。これを繰り返せば尤度関数は最終的に臨界点に達する。

Baum-Welchアルゴリズム

再推定式の導出

P と $\log P$ はHMMパラメータにより次のように表現できる。

$$P(Y, q|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(y_t)$$

$$\log P(Y, q|\lambda) = \log \pi_{q_0} + \sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log b_{qt}(y_t)$$

よって

$$Q(\lambda', \lambda) = Q_\pi(\lambda', \pi) + \sum_{i=1}^{N} Q_{a_i}(\lambda', a_i) + \sum_{i=1}^{N} Q_{b_i}(\lambda', b_i)$$

そして

$$\begin{aligned}
Q_\pi(\lambda', \pi) &= \sum_{i=1}^{N} P(Y, q_0 = i|\lambda') log \pi_i \\
Q_{a_i}(\lambda', a_i) &= \sum_{j=1}^{N} \sum_{t=1}^{T} P(Y, q_{t-1} = i, q_t = j|\lambda') \log a_{ij} \\
Q_{b_i}(\lambda', b_i) &= \sum_{t=1}^{T} P(Y, q_t = i|\lambda') \log b_i(y_t)
\end{aligned}$$

Baum-Welchアルゴリズム

$Q(\lambda', \lambda)$ は3つの独立な項に分かれているので次の制約の下でそれぞれの項を最大化することで $Q(\lambda', \lambda)$ を λ に関して最大化できる。

$$\sum_{j=1}^{N} \pi_i = 1$$

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i$$

$$\sum_{k=1}^{K} b_i(k) = 1, \forall i$$

これらは全て

$$\sum_{j=1}^{N} w_j \log y_j$$

の形をしている。これは $\sum_{j=1}^{N} y_j = 1, (y_j \geq 0)$ の制約が存在する $\{y_j\}_{j=1}^{N}$ の関数として次のような極大値に至る。

$$y_j = \frac{w_j}{\sum_{i=1}^{N} w_i}, \quad j = 1, 2, \ldots, N$$

Baum-Welchアルゴリズム

以上より、最大化処理は再推定モデル$\overline{\lambda} = [\overline{\pi}, \overline{A}, \overline{B}]$を次の点に導く。

$$\overline{\pi_i} = \frac{P(Y, q_0 = i|\lambda)}{P(Y|\lambda)}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T} P(Y, q_{t-1} = i, q_t = j|\lambda)}{\sum_{t=1}^{T} P(Y, q_{t-1} = i|\lambda)}$$

$$\overline{b_i(k)} = \frac{\sum_{t=1}^{T} P(Y, q_t = i|\lambda)\delta(y_t, v_k)}{\sum_{t=1}^{T} P(Y, q_t = i|\lambda)}$$

ここで、

$$\delta(y_t, v_k) = \begin{cases} 1 & (y_t = v_k) \\ 0 & (\text{otherwise}) \end{cases}$$

Baum-Welchアルゴリズム

前向き変数αと後向き変数βを使うと

$$
\begin{aligned}
P(Y, q_t = i|\lambda) &= \alpha_t(i)\beta_t(i) \\
P(Y|\lambda) &= \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) = \sum_{i=1}^{N} \alpha_T(i) \\
P(Y, q_{t-1} = i, q_t = j|\lambda) &= \alpha_{t-1}(i)a_{ij}b_j(y_t)\beta_t(j)
\end{aligned}
$$

これらより、前ページの式はそれぞれ(2),(4),(6)式であらわされる。

証明

$\log Z \leq Z - 1$ **の関係より、**

$$
\begin{aligned}
Q(\lambda', \lambda) - Q(\lambda', \lambda') &= \sum_q P(Y, q|\lambda') \log P(Y, q|\lambda) - \sum_q P(Y, q|\lambda') \log P(Y, q|\lambda') \\
&= \sum_q P(Y, q|\lambda') \log \frac{P(Y, q|\lambda)}{P(Y, q|\lambda')} \\
&\leq P(Y, q|\lambda') \left\{ \frac{P(Y, q|\lambda)}{P(Y, q|\lambda')} - 1 \right\} \\
&= \sum_q \{ P(Y, q|\lambda) - P(Y, q|\lambda') \} \\
&= P(Y|\lambda) - P(Y|\lambda')
\end{aligned}
$$

以上より

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow P(Y|\lambda) \geq P(Y|\lambda')$$