Învățarea din observații - Amintire

Problema Estimării (Antrenării) Modelului

Dându-se un set de secvențe observate $\mathcal{O} = [O_1 O_2 \cdots O_L]$, cum ajustăm parametrii $\lambda = (A, B, \Pi)$ ai unui MMA care încearcă să explice cel mai bine acele observații?

Secvențele de observații folosite pentru ajustarea parametrilor modelului se numesc secvențe de antrenare.
Problema antrenării este esențiala - ea permite crearea celor mai bune modele pentru fenomene reale.

Învătarea din observații - Abordare

Învățarea din observații - Abordare

Problemă

Nu se cunoaște o metodă analitică de căutare a parametrilor modelului care *maximizează* probabilitatea secvențelor observate.

Învătarea din observații - Abordare

Problemă

Nu se cunoaște o metodă analitică de căutare a parametrilor modelului care *maximizează* probabilitatea secvențelor observate.

Soluție

Putem totuși găsi $\lambda = (A, B, \Pi)$, astfel încât $\max_\lambda P(O|\lambda)$ corespunde unui maxim local, utilizând o procedură iterativă precum *algoritmul Baum-Welch*. Această metodă este o instanță a *algoritmului EM (Expectation Maximization) [DLR77]* pentru cazul MMA.

Algoritmul Baum-Welch (I)

Procedura în descriere conceptuală:

1. Avem MMA $\lambda = (A, B, \Pi)$ și o secvență observată O

Algoritmul Baum-Welch (I)

Procedura în descriere conceptuală:

1. Avem MMA $\lambda = (A, B, \Pi)$ și o secvență observată O
2. Calculăm folosindu-ne de parametrii $\alpha_t(i)$ și $\beta_t(i)$
   - nr. estimat de tranziții din S_i, pentru fiecare $1 \leq i \leq N$
   - nr. estimat de tranziții din S_i la S_j, pentru fiecare $1 \leq i \leq N, 1 \leq j \leq N$
   - nr. estimat de vizite în S_j observând simbolul v_k, pentru fiecare $1 \leq j \leq N, 1 \leq k \leq M$

Algoritmul Baum-Welch (I)

Procedura în descriere conceptuală:

1. Avem MMA $\lambda = (A, B, \Pi)$ și o secvență observată O
2. Calculăm folosindu-ne de parametrii $\alpha_t(i)$ și $\beta_t(i)$
   - nr. estimat de tranziții din S_i, pentru fiecare $1 \leq i \leq N$
   - nr. estimat de tranziții din S_i la S_j, pentru fiecare $1 \leq i \leq N, 1 \leq j \leq N$
   - nr. estimat de vizite în S_j observând simbolul v_k, pentru fiecare $1 \leq j \leq N, 1 \leq k \leq M$
3. Daca modelul este corect ne așteptăm ca
   (a) Π_i=nr. estimat de vizite în starea S_i la momentul $(t=1) = \bar{\Pi}_i$

   (b) $a_{i,j} = \frac{\text{nr. estimat de tranziții din } s_i \text{ la } s_j}{\text{nr. estimat de tranziții din } s_i} = \bar{a}_{i,j}$

   (c) $b_{j,k} = \frac{\text{nr. estimat de vizite în } s_j \text{ observând simbolul } v_k}{\text{nr. estimat de vizite în } s_j} = \bar{b}_{j,k}$

Algoritmul Baum-Welch (I)

Procedura în descriere conceptuală:

1. Avem MMA $\lambda = (A, B, \Pi)$ și o secvență observată O
2. Calculăm folosindu-ne de parametrii $\alpha_t(i)$ și $\beta_t(i)$
   - nr. estimat de tranziții din S_i, pentru fiecare $1 \leq i \leq N$
   - nr. estimat de tranziții din S_i la S_j, pentru fiecare $1 \leq i \leq N, 1 \leq j \leq N$
   - nr. estimat de vizite în S_j observând simbolul v_k, pentru fiecare $1 \leq j \leq N, 1 \leq k \leq M$
3. Daca modelul este corect ne așteptăm ca
   (a) Π_i=nr. estimat de vizite în starea S_i la momentul $(t=1) = \bar{\Pi}_i$

   (b) $a_{i,j} = \dfrac{\text{nr. estimat de tranziții din } s_i \text{ la } s_j}{\text{nr. estimat de tranziții din } s_i} = \bar{a_{i,j}}$

   (c) $b_{j,k} = \dfrac{\text{nr. estimat de vizite în } s_j \text{ observând simbolul } v_k}{\text{nr. estimat de vizite în } s_j} = \bar{b_{j,k}}$
4. Vedem că raporturile calculate din presupunerea noastră explică mai bine observația decât parametrii anteriori, i.e. $P(O|\bar{\lambda}) > P(O|\lambda)$

Algoritmul Baum-Welch (I)

Procedura în descriere conceptuală:

1. Avem MMA $\lambda = (A, B, \Pi)$ și o secvență observată O
2. Calculăm folosindu-ne de parametrii $\alpha_t(i)$ și $\beta_t(i)$
   - nr. estimat de tranziții din S_i, pentru fiecare $1 \leq i \leq N$
   - nr. estimat de tranziții din S_i la S_j, pentru fiecare $1 \leq i \leq N, 1 \leq j \leq N$
   - nr. estimat de vizite în S_j observând simbolul v_k, pentru fiecare $1 \leq j \leq N, 1 \leq k \leq M$
3. Daca modelul este corect ne așteptăm ca
   (a) Π_i=nr. estimat de vizite în starea S_i la momentul $(t=1) = \bar{\Pi}_i$
   (b) $a_{i,j} = \frac{\text{nr. estimat de tranziții din } s_i \text{ la } s_j}{\text{nr. estimat de tranziții din } s_i} = \bar{a_{i,j}}$
   (c) $b_{j,k} = \frac{\text{nr. estimat de vizite în } s_j \text{ observând simbolul } v_k}{\text{nr. estimat de vizite în } s_j} = \bar{b_{j,k}}$
4. Vedem că raporturile calculate din presupunerea noastră explică mai bine observația decât parametrii anteriori, i.e. $P(O|\bar{\lambda}) > P(O|\lambda)$
5. Atunci repetăm procesul până ce suntem mulțumiți (convergență: $P(O|\bar{\lambda}) - P(O|\lambda) \leq \epsilon$)

Algoritmul Baum-Welch (II)

Algoritmul Baum-Welch (II)

Definim întâi niște variabile auxiliare:

$\xi_{t,i,j} = \xi_t(i,j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda)$
Probabilitatea de a fi în starea s_i la momentul t **și** în starea s_j la momentul $t+1$, condiționat de parametrii modelului curent și secvența observată.

Algoritmul Baum-Welch (II)

Definim întâi niște variabile auxiliare:

$\xi_{t,i,j} = \xi_t(i,j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda)$
Probabilitatea de a fi în starea s_i la momentul t **și** în starea s_j la momentul $t+1$, condiționat de parametrii modelului curent și secvența observată.

$\gamma_{t,i} = \gamma_t(i) = P(q_t = s_i | O, \lambda)$
Probabilitatea de a fi în starea s_i la momentul t, condiționat de parametrii modelului curent și secvența observată.

Algoritmul Baum-Welch (II)

Definim întâi niște variabile auxiliare:

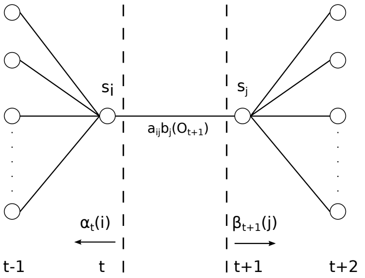
$\xi_{t,i,j} = \xi_t(i,j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda)$
Probabilitatea de a fi în starea s_i la momentul t **și** în starea s_j la momentul $t+1$, condiționat de parametrii modelului curent și secvența observată.

$\gamma_{t,i} = \gamma_t(i) = P(q_t = s_i | O, \lambda)$
Probabilitatea de a fi în starea s_i la momentul t, condiționat de parametrii modelului curent și secvența observată.

Din definiții rezultă că:

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

Algoritmul Baum-Welch (III)



Secvența de operații necesară pentru calculul evenimentului mixt ca sistemul se află în starea S_i la momentul t și în starea S_j la momentul $t+1$

$$\alpha_{t,i} = P(o_1, o_2, \ldots, o_t, q_t = S_i|\lambda) \quad (36)$$

$$\beta_{t,i} = P(o_{t+1} o_{t+2} \cdots o_T | q_t = S_i, \lambda) \quad (37)$$

$$\xi_t(i,j) = \frac{\alpha_{t,i} \cdot a_{i,j} \cdot b_j(o_{t+1}) \cdot \beta_{t+1,j}}{P(O|\lambda)} = \frac{\alpha_{t,i} \cdot a_{i,j} \cdot b_j(o_{t+1}) \cdot \beta_{t+1,j}}{\sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_{t,k} \cdot a_{k,l} \cdot b_l(o_{t+1}) \cdot \beta_{t+1,l}} \quad (38)$$

Algoritmul Baum-Welch (IV)

Cum ne ajută aceste variabile auxiliare?

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{numărul estimat de tranziții din } S_i$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{numărul estimat de tranziții din } S_i \text{ la } S_j$$

Algoritmul Baum-Welch (V)

$$\bar{\pi}_i = \text{ nr. estimat de vizite în starea } S_i \text{ la momentul } (t = 1) = \gamma_1(i) \tag{39}$$

Algoritmul Baum-Welch (V)

$$\bar{\pi}_i = \text{ nr. estimat de vizite în starea } S_i \text{ la momentul } (t = 1) = \gamma_1(i) \tag{39}$$

$$\begin{aligned} \bar{a_{i,j}} &= \frac{\text{nr. estimat de tranziții din } S_i \text{ la } S_j}{\text{nr. estimat de tranziții din } S_i} \\ &= \frac{\displaystyle\sum_{t=1}^{T-1} \xi_t(i,j)}{\displaystyle\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \tag{40}$$

Algoritmul Baum-Welch (V)

$$\bar{\pi_i} = \text{ nr. estimat de vizite în starea } S_i \text{ la momentul } (t=1) = \gamma_1(i) \tag{39}$$

$$\begin{aligned} \bar{a_{i,j}} &= \frac{\text{nr. estimat de tranziții din } S_i \text{ la } S_j}{\text{nr. estimat de tranziții din } S_i} \\ &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \tag{40}$$

$$\begin{aligned} \bar{b_{j,k}} &= \frac{\text{nr. estimat de vizite în } S_j \text{ observând simbolul } v_k}{\text{nr. estimat de vizite în } S_j} \\ &= \frac{\sum_{t=1, O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \end{aligned} \tag{41}$$

Algoritmul Baum-Welch (VI)

Algoritmul 5 Algoritm Baum-Welch

intrări: $O \leftarrow$ secvența de observații, $\epsilon \leftarrow$ prag de convergență

{*Initializare*}

init. uniformă Π $(\Pi_i = 1/N, 1 \leq i \leq N)$

init. aleatoare $a_{i,j}$, a. î. $\sum_{j=1}^{N} a_{i,j} = 1, \forall i = \overline{1,N}$

init. uniformă $b_{j,k}$ $(b_{j,k} = 1/M, 1 \leq j \leq N, 1 \leq k \leq M)$

init $log(P(O|\bar{\lambda})) = 0$

repeat

$\quad log(P(O|\lambda)) = log(P(O|\bar{\lambda}))$

$\quad$ {*E STEP - calculeaza variantele scalate pentru* α *și* β *și probabilitatea curentă (log likelihood -* $log(P(O|\bar{\lambda}))$*) a secvenței observate*}

$\quad [log(P(O|\bar{\lambda})), \hat{\alpha}, \hat{\beta}, Scale] = forward\_backward(O, \Pi, A, B)$

$\quad$ {M STEP - recalculeaza Π, A și B}

$\quad \Pi = update\_pi\_procedure(\hat{\alpha}, \hat{\beta}, Scale)$

$\quad A = update\_A\_procedure(O, \hat{\alpha}, \hat{\beta}, Scale)$

$\quad B = update\_B\_procedure(O, \hat{\alpha}, \hat{\beta}, Scale)$

until $log(P(O|\bar{\lambda})) - log(P(O|\lambda)) < \epsilon$

Algoritmul Baum-Welch (VI)

Algoritmul 6 Algoritm Baum-Welch

Function update_pi_procedure($\hat{\alpha}$, $\hat{\beta}$, *Scale*)
for $i = 1$ to N **do**
$\quad \Pi_i = \frac{\hat{\alpha}_1(i) \cdot \hat{\beta}_1(i) / Scale(1)}{\sum_{j=1}^{N} \hat{\alpha}_1(j) \cdot \hat{\beta}_1(j) / Scale(1)}$
end for
return Π
EndFunction

Function update_A_procedure(O, $\hat{\alpha}$, $\hat{\beta}$, *Scale*)
for $i = 1$ to N **do**
$\quad$ **for** $j = 1$ to N **do**
$\quad\quad a_{i,j} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_{t,i} \cdot a_{ij} \cdot b_j(o_{t+1}) \cdot \hat{\beta}_{t+1,j}}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} \hat{\alpha}_{t,i} \cdot a_{i,j} \cdot b_j(o_{t+1}) \cdot \hat{\beta}_{t+1,j}}$
$\quad$ **end for**
end for
return a
EndFunction

Algoritmul Baum-Welch (VI)

Algoritmul 7 Algoritm Baum-Welch

Function update_B_procedure(O, $\hat{\alpha}$, $\hat{\beta}$, *Scale*)
for $j = 1$ to N **do**
    for $k = 1$ to M **do**
        $b_{j,k} = \frac{\sum_{t=1, O(t)=v_k}^{T} \hat{\alpha}_t(j) \cdot \hat{\beta}_t(j) / Scale(t)}{\sum_{t=1}^{T} \hat{\alpha}_t(j) \cdot \hat{\beta}_t(j) / Scale(t)}$
    end for
end for
return b
EndFunction