Spis rzeczy

5 Uliniowienie wielu sekwencji **77**
5.8 Profilowe ukryte modele Markowa 77

5 Uliniowienie wielu sekwencji

5.8 Profilowe ukryte modele Markowa

Na zakończenie omawiania problematyki uliniowienia wielu sekwencji zajmiemy się wykorzystaniem ukrytych modeli Markowa. Niech $\mathcal{P}$ będzie profilem składającym się z m rozkładów prawdopodobieństw $b_1, \ldots, b_m$, każde na alfabecie $\Sigma \cup \{-\}$ (por. Sekcja 5.1). Dla $x \in \Sigma \cup \{-\}$, $b_i(x)$ jest prawopodobieństwem pojawienia sie symbolu x w i-tej kolumnie (pewnego) uliniowienia, którego reprezentacją jest $\mathcal{P}$. Ponadto przyjmujemy prawdopodobieństwo b_- na zbiorze $\Sigma \cup \{-\}$, tzw. *prawdopodobieństwo tła*. Jest to prawdopodobieństwo $b_-(x)$ pojawienia się (wstawienia) symbolu $x \in \Sigma$ z zewnątrz. Przyjmujemy $b_-(-) = 0$. Mając daną funkcję podobieństwa s na Σ (określoną również dla $-$) oraz pewne uliniowienie $\mathcal{P}^\#, S^\#$, profilu $\mathcal{P}$ ze słowem S, wartość tego uliniowienia definiuje się jako

$$\sum_{i=1}^{n} s(S^\#(i), \mathcal{P}^\#(i)),$$

gdzie $n = |S^\#| = |\mathcal{P}^\#|$ oraz dla $x \in \Sigma \cup \{-\}$ i dla $\xi \in \{1, \ldots, n\} \cup \{-\}$,

$$s(x, \xi) = \sum_{y \in \Sigma \cup \{-\}} s(x, y) \cdot b_\xi(y).$$

Zdefiniujemy pewien ukryty model Markowa $M_\mathcal{P}$ związany z profilem $\mathcal{P}$.

Topologia połączeń $M_\mathcal{P}$

Model $M_\mathcal{P}$ ma następujące rodzaje stanów:

- `stt`– stan początkowy
- `stp`– stan końcowy

- $d_1, \ldots, d_m$ – stany reprezentujące *dopasowanie*; będziemy je nazywać *d-stanami.*
- $u_1, \ldots, u_m$ – stany reprezentujące *usunięcie*; będziemy je nazywać *u-stanami.*
- $w_0, w_1, \ldots, w_m$ – stany reprezentujące *wstawienie*; będziemy je nazywać *w-stanami.*

Stany te są połaczone w następujący sposób (por. rysunek).

Prawdopodobieństwa emisji

Stany: `stt`, `stp`, $u_1, \ldots, u_m$ nie emitują żadnych symboli. W stanie w_i symbol $x \in \Sigma$ jest emitowany z prawdopodobieństwem $b_-(x)$. W stanie d_i symbol $x \in \Sigma \cup \{-\}$ jest emitowany z prawdopodobieństwem $b_i(x)$.

Sposób emitowania pewnego słowa S przez $M_{\mathcal{P}}$ jest związany z pewnym uliniowieniem S z profilem $\mathcal{P}$. Wyemitowanie symbolu x w stanie w_i oznacza, że w uliniowieniu, o którym mowa powyżej, symbol x stoi naprzeciw spacji (spacja ta jest wstawiona pomiędzy kolumnami i oraz $i+1$, dla $0 < i < m$; dla skrajnych wartości i, spacja jest wstawiona na skrajnej pozycji). Natomiast wyemitowanie symbolu x w stanie d_i oznacza, że x jest uliniowione z kolumną i profilu $\mathcal{P}$. Wreszcie znalezienie się układu w stanie u_i oznacza, że spacja w rozszerzeniu S jest uliniowiona z i-tą kolumną profilu.

Przykład 5.8.1 Dla przykładu przyjmijmy, że $m = 5$ oraz, że słowo *ssbbc* zostało wyemitowane poprzez następujący ciąg stanów

$$\texttt{stt},\ (d_1, a),\ (w_1, a),\ (d_2, b),\ u_3,\ (d_4, b),\ (d_5, c),\ \texttt{stp}$$

Odpowiada to uliniowieniu

```
aab-bc
1-2345
```

□

Uliniowienie słowa z profilem (algorytm Viterbiego)

Przyjmijmy, że mamy model $M_{\mathcal{P}}$, w którym dla każdej pary stanów s, s', jeśli w $M_{\mathcal{P}}$ istnieje strzałka od s do s', to określone jest prawdopodobieństwo przejścia $p_{s,s'}$.

Niech $S \in \Sigma^*$ i niech $\Pi = s_0, \ldots, s_n$ będzie drogą w grafie połączeń $M_{\mathcal{P}}$ taką, że $s_0 = \texttt{stt}$ oraz $s_n = \texttt{stp}$. Niech Q_Π będzie zbiorem wszystkich d-stanów oraz w-stanów występujących w Π. Zakładając, że zmiana stanów odbywała się wzdłóż drogi Π, prawdopodobieństwo wyemitowania S wynosi

$$P(S, \Pi) = \prod_{i=1}^{n} p_{s_{i-1}, s_i} \cdot f_{s_i}(S^{\#}(i)),$$

gdzie

$$S^{\#}(i) = \begin{cases} - & \text{gdy } s_i \notin Q_\Pi, \\ S(j) & \text{gdy } s_i \in Q_\Pi \text{ oraz } j = |\{s_l \in Q_\Pi \mid l \leq i\}|. \end{cases}$$

Ponadto
$f_{d_i} = b_i$
$f_{w_i} = b_-$
$f_\xi(-) = 1$, dla $\xi \notin Q_\Pi$.

Tak więc zadanie uliniowienia słowa S z profilem $\mathcal{P}$ polega na znalezieniu maksimum funkcji $P(S, \Pi)$, po wszystkich drogach Π spełniających wyżej wymienione warunki. Zastosujemy algorytm Viterbiego (por. Sekcja 4.1) znajdowania optymalnej drogi Π.

Dla $\xi \in \{d_1, \ldots, d_m\} \cup \{u_1, \ldots, u_m\} \cup \{w_0, \ldots, w_m\} \cup \{\texttt{stt}\}$ oraz $0 \leq i \leq |S|$, niech $V_\xi(i)$ będzie maksymalnym prawdopodobieństwem wyemitowania $S[1..i]$ przez $M_{\mathcal{P}}$ przy przebiegu pewnej drogi kończącej się stanem ξ tak, że symbol $S(i)$ (o ile $i > 0$ oraz ξ jest d-stanem lub u-stanem) jest wyemitowany w stanie ξ.

Mamy wówczas następujące równania (przyjmujemy, że $d_0 = \texttt{stt}$ oraz $p_{u_0,u_1} = p_{u_0,d_1} = 0$). Najpierw warunki początkowe

$$V_{\texttt{stt}}(i) = \begin{cases} 1 & \text{gdy } i = 0, \\ 0 & \text{gdy } i > 0. \end{cases}$$

$$V_{d_j}(0) = V_{w_k}(0) = 0,$$

dla $j = 1, \ldots, m$ oraz $k = 0, \ldots m$.
Dla $i > 0$ oraz $j > 0$ mamy

$$V_{d_j}(i) = b_j(S(i)) \cdot \max[V_{d_{j-1}}(i-1) \cdot p_{d_{j-1},d_j},\ V_{w_{j-1}}(i-1) \cdot p_{w_{j-1},d_j},\ V_{u_{j-1}}(i-1) \cdot p_{u_{j-1},d_j}].$$

Dla $i \geq 0$ oraz $j > 0$ mamy

$$V_{u_j}(i) = \max[V_{d_{j-1}}(i) \cdot p_{d_{j-1},u_j},\ V_{w_{j-1}}(i) \cdot p_{w_{j-1},u_j},\ V_{u_{j-1}}(i) \cdot p_{u_{j-1},u_j}].$$

Wreszcie, dla $i > 0$ oraz $j > 0$ mamy

$$V_{w_j}(i) = b_{-}(S(i)) \cdot \max[V_{d_j}(i-1) \cdot p_{d_j,w_j},\ V_{w_j}(i-1) \cdot p_{w_j,w_j},\ V_{u_j}(i-1) \cdot p_{u_j,w_j}].$$

Mając wyliczone funkcje V_ξ, optymalną wartość obliczamy następująco:

$$P(S) = \max[V_{d_m}(|S|) \cdot p_{d_m,\mathtt{stp}}, V_{u_m}(|S|) \cdot p_{u_m,\mathtt{stp}}, V_{w_m}(|S|) \cdot p_{w_m,\mathtt{stp}}].$$

Oczywiście złożoność powyższego algorytmu wynosi (czas i przestrzeń) $O(|S|m)$.

Gdy prawdopodobieństwa przejść modelu $M_\mathcal{P}$ nie są dane, to stosując algorytm Bauma-Welscha (por. Sekcja 4.3.1) możemy dobrać poszukiwane prawdopodobieństwa dla zadanych słów $S_1, \ldots, S_k$ tak, aby zmaksymalizować wartość $\prod_{i=1}^{k} P(S_i)$. W tym celu wyliczamy najpierw prawdopodobieństwa prefiksowe i sufiksowe (por. Sekcje 4.2.1 i 4.2.2). Szczegóły pomijamy, gdyż nie różnią się one znacznie od szczegółów opisanych w w/w Sekcjach[1]

Ukryte modele Markowa a uliniowienie wielu sekwencji

Dane sekwencje $S_1, \ldots, S_k$. Jeśli mamy również profil $\mathcal{P}$, to budujemy model $M_\mathcal{P}$, dla którego dobieramy wartości prawdopodobieństw przejść, stosując algorytm Bauma-Welscha. Następnie dla każdego słowa S_i stosujemy algorytm Viterbiego, obliczając optymalne uliniowienie słowa S_i z profilem $\mathcal{P}$. Ta matoda daje nam pewne uliniowienie słów $S_1, \ldots, S_k$ z profilem $\mathcal{P}$.

Jeśli natomiast nie mamy profilu, to wybieramy długość (przyszłego profilu) i stosujemy algorytm Bauma-Welcha, dobierając nie tylko prawdopodobieństwa przejść, ale i prawdopodobieństwa emisji (stosują się tutaj zastrzeżenia, że w stanach $\mathtt{stt}, \mathtt{stp}, u_1, \ldots, u_m$ żaden symbol nie jest emitowany). To daje nam model $M_\mathcal{P}$. Dalej, stosujemy algorytm Viterbiego do każdego ze słów S_i i modelu $M_\mathcal{P}$. Ta metoda daje nam pewien profil $\mathcal{P}$ oraz uliniowienie $S_1, \ldots, S_k$ z profilem $\mathcal{P}$.

[1] Różnice są podobnej natury co różnice pomiędzy algorytmem Viterbiego opisanym w Sekcji 4.1, a algorytmem opisanym powyżej.