

Processament del Llenguatge Humà (GIA-PLH)

Exercicis sobre atributs per aprendre models CRF de reconeixement i classificació de seqüències de paraules

SOLUCIONS

Recordemos los modelos linear-chain CRF para la predicción de secuencias y pensemos en la tarea de NERC. Un modelo CRF bigrama computa:

$$\text{tags}(x_{1:n}) = \operatorname{argmax}_{y_{1:n} \in \mathcal{Y}^n} \bar{\lambda}_{1:k} \cdot \mathbf{f}(x_{1:n}, y_{i:n}) \quad (1)$$

$$= \operatorname{argmax}_{y_{1:n} \in \mathcal{Y}^n} \sum_{i=1}^n \bar{\lambda}_{1:k} \cdot \mathbf{f}(x_{1:n}, i, y_{i-1}, y_i) \quad (2)$$

donde:

$x_{1:n}$: secuencia de entrada de n tokens (x_i es el i -ésimo token)

$y_{1:n}$: secuencia de salida de n etiquetas (\mathcal{Y} es el conjunto de etiquetas válidas)

$\mathbf{f}(x_{1:n}, i, y_{i-1}, y_i)$: función que devuelve un vector definido en $\{0, 1\}^k$, resultante de aplicar las k funciones de atributos $\bar{\mathbf{f}}_{1:k}$, relativas al bigrama y_{i-1}, y_i en la posición i de la secuencia de entrada (se asume que y_0 es una etiqueta especial START que indica el inicio de secuencia).

$\mathbf{f}(x_{1:n}, y_{1:n})$: función que devuelve el vector global de atributos definido en \mathbb{N}^k , resultante de aplicar las k funciones de atributos $\bar{\mathbf{f}}_{1:k}$ a toda la secuencia $x_{1:n}$.

$\bar{\lambda}_{1:k}$: vector de parámetros del modelo.

Ejercicio 1.

Especificamos los atributos f_i usando plantillas. Por ejemplo, la siguiente plantilla captura la palabra y la etiqueta actuales:

$$\mathbf{f}_{1,l,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } x_i = a \text{ and } y_i = l \\ 0 & \text{otherwise} \end{cases}$$

Escribe las plantillas correctas de atributos que capturen los siguientes patrones. Justifica tus respuestas si es necesario

- $\mathbf{f}_{2,a}$: la palabra actual es la primera de la secuencia, empieza con mayúscula y su etiqueta es a
- $\mathbf{f}_{3,s,a}$: el prefijo de 3 letras de la palabra actual es s y la etiqueta actual es a
- $\mathbf{f}_{4,w,a,b}$: en orden, la palabra actual, la etiqueta actual y la etiqueta anterior
- $\mathbf{f}_{5,w,v,a}$: en orden, las 2 palabras anteriores y la etiqueta actual
- $\mathbf{f}_{6,a,b,c}$: en orden, las dos etiquetas previas y la etiqueta actual

SOLUCIÓN

$$\mathbf{f}_{2,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } i = 1 \text{ and } \text{capitalized}(x_i) \text{ and } y_i = a \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{3,s,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } \text{prefix}(x_i) = s \text{ and } y_i = a \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{4,w,a,b}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } x_i = w \text{ and } y_{i-1} = a \text{ and } y_i = b \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{5,w,v,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } x_{i-2} = w \text{ and } x_{i-1} = v \text{ and } y_i = a \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{5,a,b,c}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } y_{i-2} = a \text{ and } y_{i-1} = b \text{ and } y_i = c \\ 0 & \text{otherwise} \end{cases}$$

Ejercicio 2.

1. Dado el siguiente ejemplo de entrenamiento the/DT dog/NN saw/VBD the/DT man/NN, si lo convertimos en pares (x, y) , donde $x = (y_{i-2}, y_{i-1}, O, i)$ para entrenar un CRF basado en trigramas para *PoS tagging*, ¿cuáles de los siguientes pares están en el conjunto de entrenamiento?

- (a) $x = (\text{DT}, \text{NN}, \text{the dog saw the man}, 3); y = \text{NN}$
- (b) $x = (\text{VBD}, \text{DT}, \text{the dog saw the man}, 3); y = \text{VBD}$
- (c) $x = (\text{DT}, \text{NN}, \text{the dog saw the man}, 3); y = \text{VBD}$
- (d) $x = (\text{DT}, \text{NN}, \text{the dog saw the man}, 4); y = \text{NN}$

2. Lista todos los pares (x, y) que pueden ser generados a partir del conjunto de entrenamiento. Supón que tenemos, además, los estados virtuales $y_{-2} = y_{-1} = \text{START}$.

SOLUCIÓN

- 1. (a) no se cumple porque la posición 3 tiene $y = \text{VBD}$ en vez de NN .
- (b) no se cumple, dado que las etiquetas anteriores para la posición 3 son $y_{i-2} = \text{DT}$ y $y_{i-1} = \text{NN}$ en vez de VBD y DT , respectivamente.
- (c) sí se cumple, ya que la etiqueta para la posición 3 es $y = \text{VBD}$ y las etiquetas anteriores son $y_{i-2} = \text{DT}$ y $y_{i-1} = \text{NN}$
- (d) no se cumple, puesto que la posición 4 tiene $y = \text{DT}$ en vez de NN .

2. Los pares generados son:

$$x = (\text{START}, \text{START}, \text{the dog saw the man}, 1); y = \text{DT}$$

$$x = (\text{START}, \text{DT}, \text{the dog saw the man}, 2); y = \text{NN}$$

$$x = (\text{DT}, \text{NN}, \text{the dog saw the man}, 3); y = \text{VBD}$$

$$x = (\text{NN}, \text{VBD}, \text{the dog saw the man}, 4); y = \text{DT}$$

$$x = (\text{VBD}, \text{DT}, \text{the dog saw the man}, 5); y = \text{NN}$$

Ejercicio 3.

Queremos tratar la tarea de *PoS tagging* con un modelo de CRFs basado en bigramas que compute la etiqueta para cada palabra de la siguiente manera:

$$\text{tag}(x_{1:n}, i) = \underset{y_i \in \mathcal{Y}}{\operatorname{argmax}} \bar{\lambda}_{1:k} \cdot \mathbf{f}(x_{1:n}, i, y_{i-1}, y_i)$$

Definimos las siguientes plantillas de atributos:

- Plantilla 1: La etiqueta actual es a :

$$f_{1,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } y_i = a \\ 0 & \text{otherwise} \end{cases}$$

- Plantilla 2: La palabra actual empieza con mayúscula y la etiqueta actual es a :

$$f_{2,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } x_i \text{ is capitalized and } y_i = a \\ 0 & \text{otherwise} \end{cases}$$

- Plantilla 3: La etiqueta anterior es a y la actual es b :

$$f_{3,a,b}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } y_{i-1} = a \text{ and } y_i = b \\ 0 & \text{otherwise} \end{cases}$$

- Propón valores del vector $\bar{\lambda}_{1:n}$ para funciones de atributos apropiados para clasificar correctamente las palabras de las siguientes oraciones. Intenta usar el mínimo número de pesos no nulos. Justifica tu selección.

x : John programs bugs
 y : E V N

x : Mary runs programs
 y : E V N

x : Mary bugs John
 y : E V E

x : programs print results
 y : N V N

SOLUCIÓN

- $w_{1,V} = 1$ will score +1 for any word to be tagged as a verb. This will solve correctly all verbs, and introduce a wrong bias in the other words.
- $w_{3,V,N} = 2$ will score +2 in favor of tag N for any word after a V. This will overcome the first feature and correctly solve the noun *results* in the last sentence.
- $w_{3,START,N} = 3$ will score +3 in favor of tag N for any word at the beginning of the sentence. This will overcome the first feature and solve correctly the noun *programs* in the last sentence. If we used +2 here, we would get a tie between combinations V-N and N-V at sentence beginning, so we use +3.
- $w_{2,E} = 4$ will score +4 for all capitalized words to be tagged E. This will overcome the previous features and solve properly all occurrences of *John* and *Mary*.

Let's apply Viterbi algorithm to check these weights work for all given sentences:

	John	programs	bugs
	Mary	runs	programs
E	E:4	E-E: 4+0=4 N-E: 3+0=3 V-E 1+0=1	E-E-E: 4+0=4 E-N-E: 4+0=4 E-V-E: 5+0=5
N	N:3	E-N: 4+0=4 N-N: 3+0=3 V-N: 1+2=3	E-E-N: 4+0=0 E-N-N: 4+0=0 E-V-N: 5+2=7
V	V: 1	E-V: 4+1=5 N-V: 3+1=4 V-V: 1+1=2	E-E-V: 4+1=5 E-N-V: 4+1=5 E-V-V: 5+1=6

←Best

For sentences *John programs bugs* and *Mary runs programs*, the best sequence is E-V-N with a score of 7, higher than any other combination.

	Mary	bugs	John
E	E: 4	E-E: 4+0=4 N-E: 3+0=3 V-E 1+0=1	E-E-E: 4+4=8 E-N-E: 4+4=8 E-V-E: 5+4=9
N	N: 3	E-N: 4+0=4 N-N: 3+0=3 V-N: 1+2=3	E-E-N: 4+0=0 E-N-N: 4+0=0 E-V-N: 5+2=7
V	V:1	E-V: 4+1=5 N-V: 3+1=4 V-V: 1+1=2	E-E-V: 4+1=5 E-N-V: 4+1=5 E-V-V: 5+1=6

←Best

For sentence *Mary bugs John*, the best sequence is E-V-E with a score of 9, higher than any other combination.

	programs	print	results
E	E: 0	E-E: 0+0=0 N-E: 3+0=3 V-E 1+0=1	N-E-E: 3+0=3 N-N-E: 3+0=3 N-V-E: 4+0=4
N	N: 3	E-N: 0+0=0 N-N: 3+0=3 V-N: 1+2=3	N-E-N: 3+0=3 N-N-N: 3+0=0 N-V-N: 4+2=6
V	V: 1	E-V: 0+1=1 N-V: 3+1=4 V-V: 1+1=2	N-E-V: 3+1=4 N-N-V: 3+1=4 N-V-V: 4+1=5

←Best

For sentence *programs print results*, the best sequence is N-V-N with a score of 6, higher than any other combination.

Ejercicio 4.

Supongamos que trabajamos en la tarea de *PoS tagging* con un CRF basado en trigramas y usando el conjunto de etiquetas {DT, V, NN, ADV, PREP}. Supongamos, además, que definimos una *historia* como $h = \langle t_{i-2}, t_{i-1}, x_{1:n}, i \rangle$.

1. ¿Cuántas posibles historias hay para una secuencia de entrada dada, $x_{1:n}$, y un valor de i fijado?
2. ¿Cuáles de las siguientes funciones de atributos son correctas?

$$f_1(h, t_i) = \begin{cases} 1 & \text{if } t_i = V \text{ and } t_{i-1} = \text{PREP} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(h, t_i) = \begin{cases} 1 & \text{if } t_i = V \text{ and } w_{i-2} = \text{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(h, t_i) = \begin{cases} 1 & \text{if } t_i = V \text{ and } t_{i-3} = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(h, t_i) = \begin{cases} 1 & \text{if } t_i = V \text{ and } t_{i+1} = \text{PREP} \text{ and } w_2 = \text{cow} \\ 0 & \text{otherwise} \end{cases}$$

3. Calcula el vector global de atributos $f(\mathcal{X}, \mathcal{Y})$ para la secuencia de entrada $\mathcal{X} = \text{the dog walked to a park}$ y la secuencia de salida $\mathcal{Y} = \text{DT NN V PREP DT NN}$, usando las siguientes funciones de atributos:

$$\mathbf{f}_1(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{NN and } w_i = \text{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{NN and } t_{i-1} = \text{DT} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{NN and } t_{i-1} = \text{DT and } w_{i-1} = \text{the} \\ 0 & \text{otherwise} \end{cases}$$

4. Dada la historia $h = (t_{i-2}, t_{i-1}, x_{1:n}, 5) = (\text{V, DT, the man saw the dog in the park}, 5)$, ¿cuáles de las siguientes funciones de atributos cumplen $f(h, \text{NN}) = 1$?

$$\mathbf{f}_1(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{NN and } w_i = \text{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{DT and } w_i = \text{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{NN and } w_{i+1} = \text{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_4(h, t_i) = \begin{cases} 1 & \text{if } t_i = \text{NN and } t_{i-1} = \text{DT} \\ 0 & \text{otherwise} \end{cases}$$

SOLUCIÓN

- Cada historia tiene la forma $h = (t_{i-2}, t_{i-1}, x_{[1:n]}, i)$. Si fijamos la secuencia $\mathcal{X} = x_{[1:n]}$ y la posición i , sólo pueden variar los parámetros t_{i-2} y t_{i-1} . Dado que a cada uno de ellos se les puede asignar cualquier etiqueta PoS del conjunto $\{\text{DT, V, NN, ADV, PREP}\}$, el número de combinaciones posibles es $5 \times 5 = 25$.
- Las funciones f_1 y f_2 son correctas porque ambas usan tanto el estado actual, t_i , como elementos de h . Dado que f_3 y f_4 usan t_{i-3} y t_{i+1} , respectivamente, y dichos elementos no están definidos en h , estas funciones son incorrectas.
- Dados \mathcal{X} e \mathcal{Y} , se cumplen las siguientes funciones de atributos:

\mathcal{X}	the	dog	walked	to	a	park
\mathcal{Y}	DT	NN	V	PREP	DT	NN
Funciones	-	f_1 f_2 f_3	-	-	-	f_2

Por lo tanto, el vector global de atributos, $f(\mathcal{X}, \mathcal{Y})$, que se obtiene es $(f_1, f_2, f_3) = (1, 2, 1)$

4. La historia h nos restringe a la posición $i = 5$, que corresponde a la palabra dog con etiqueta NN. Si evaluamos cada función sobre la secuencia $x_{1:n}$ obtenemos:

$$f_1(h, \text{NN}) = 1, \text{ dado que } t_i = \text{NN y } x_5 = \text{dog}$$

$$f_2(h, \text{NN}) = 0, \text{ dado que } t_i \neq \text{DT}$$

$$f_3(h, \text{NN}) = 0, \text{ dado que } x_6 \neq \text{dog}$$

$$f_4(h, \text{NN}) = 1, \text{ dado que } t_i = \text{NN y } t_{i-1} = \text{DT}$$

Ejercicio 5.

Nuestro cliente requiere identificar automáticamente enfermedades mencionadas en informes médicos. Para ello, queremos entrenar un CRF basado en bigramas usando, pues, el contexto $h = (t_{i-1}, o_{1:n}, i)$ mas las siguientes plantillas de atributos:

$$\begin{aligned} \mathbf{f}_{1,a}(h, t_i) &= \begin{cases} 1 & \text{if } pos(o_{i-1}) = N \text{ and } t_{i-1} = O \text{ and } t_i = a \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{f}_2(h, t_i) &= \begin{cases} 1 & \text{if } suf(o_{i-1}) = 'ing' \text{ and } t_i = B \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{f}_{3,a,b}(h, t_i) &= \begin{cases} 1 & \text{if } o_{i-1} = a \text{ and } t_i = b \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{f}_{4,a,b,c}(h, t_i) &= \begin{cases} 1 & \text{if } o_{i-1} = a \text{ and } t_i = b \text{ and } pos(o_i) = c \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{f}_{5,a,b}(h, t_i) &= \begin{cases} 1 & \text{if } o_i = a \text{ and } capitalized(o_{i-1}) \text{ and } t_i = b \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Dadas dichas plantillas y la siguiente oración de entrenamiento:

i	1	2	3	4	5	6	7	8	9	10	11	12	13
w	Fragile-X	syndrome	is	an	inherited	form	of	mental	retardation	involving	mitral	valve	prolapse
pos	N	N	V	D	JJ	N	P	JJ	N	V	JJ	N	N
t	B	I	O	O	O	O	O	O	O	O	B	I	I

Escribe las funciones de atributos que se derivarían para las siguientes palabras:

- syndrome*
- involving*
- mitral*

SOLUTION

- Funciones para *syndrome* ($i = 2$): $(f_{3, \text{Fragile-X, I}}, f_{4, \text{Fragile-X, I, N}}, f_{5, \text{syndrome, I}})$
- Funciones para *involving* ($i = 10$): $(f_{1, O}, f_{3, \text{retardation, O}}, f_{4, \text{retardation, O, V}})$
- Funciones para *mitral* ($i = 11$): $(f_2, f_{3, \text{involving, B}}, f_{4, \text{involving, B, JJ}})$