Processament del Llenguatge Humà (GIA-PLH)

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

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

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

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

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* $\overline{\mathbf{f}}_{1:k}$*, 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* $\overline{\mathbf{f}}_{1:k}$ *a toda la secuencia* $x_{1:n}$*.*

$\overline{\lambda}_{1:k}$*: 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

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

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

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

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

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

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

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

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:

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

1. Propón valores del vector $\overline{\lambda}_{1:n}$ 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

Ejercicio 3.

Supongamos que trabajamos en la tarea de *PoS tagging* con un CRF basado en trigramas y usando el conjunto de etiquetas $\{\texttt{DT}, \texttt{V}, \texttt{NN}, \texttt{ADV}, \texttt{PREP}\}$. 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?

$$\mathbf{f}_1(h, t_i) = \begin{cases} 1 & \text{if } t_i = \texttt{V} \text{ and } t_{i-1} = \texttt{PREP} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h, t_i) = \begin{cases} 1 & \text{if } t_i = \texttt{V} \text{ and } w_{i-2} = \texttt{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h, t_i) = \begin{cases} 1 & \text{if } t_i = \texttt{V} \text{ and } t_{i-3} = \texttt{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_4(h, t_i) = \begin{cases} 1 & \text{if } t_i = \texttt{V} \text{ and } t_{i+1} = \texttt{PREP} \text{ and } w_2 = \texttt{cow} \\ 0 & \text{otherwise} \end{cases}$$

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

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

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

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

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

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

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

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

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

Ejercicio 4.

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

$$\mathbf{f}_{1,a}(h, t_i) = \begin{cases} 1 & \text{if } pos_{i-1} = \text{N} \text{ and } t_{i-1} = \text{O} \text{ and } t_i = a \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h, t_i) = \begin{cases} 1 & \text{if } suf(w_{i-1}) = \text{'ing'} \text{ and } t_i = \text{B} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{3,a,b}(h, t_i) = \begin{cases} 1 & \text{if } w_{i-1} = a \text{ and } t_i = b \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{4,a,b,c}(h, t_i) = \begin{cases} 1 & \text{if } w_{i-1} = a \text{ and } t_i = b \text{ and } pos_i = c \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{5,a,b}(h, t_i) = \begin{cases} 1 & \text{if } w_i = a \text{ and } capitalized(w_{i-1}) \text{ and } t_i = b \\ 0 & \text{otherwise} \end{cases}$$

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:

a) *syndrome*

b) *involving*

c) *mitral*