Mining Unstructured Data
Exercises on features for learning CRF models to classify word sequences

Features for bigram-factored sequence annotation models

Recall the linear CRF models for sequence prediction, and think of a named entity task. A bigram CRF model computes:

$$tags(x_{1:n}) = \operatorname*{argmax}_{y_{1:n} \in \mathcal{Y}^n} \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(x, i, y_{i-1}, y_i) \qquad (1)$$

where $x_{1:n}$ is an input sentence of n tokens (x_i is the i-th token), $y_{1:n}$ is an output sequence of n tags ($\mathcal{Y}$ is the set of valid tags). $\mathbf{f}(x, i, y_{i-1}, y_i)$ is a function returning a feature vector of the bigram y_{i-1}, y_i at position i of the sentence (assume that y_0 is a special tag START *that indicates the start of the sequence). $\mathbf{w}$ is a vector of parameters of the same dimensionality of the feature vectors.*

Exercise 1.

We specify features using templates. For example, the following template captures the current word and the current tag:

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

Write feature templates that capture the following patterns. Justify your answers if necessary.

- $\mathbf{f}_{2,a}$: the current word is the first of the sentence, it is capitalized, and its tag is a
- $\mathbf{f}_{3,s,a}$: 3-letter prefix of the current word, together with the current tag
- $\mathbf{f}_{4,w,a,b}$: the current word, the current tag, and the previous tag
- $\mathbf{f}_{5,w,v,a}$: the two previous words and the current tag
- $\mathbf{f}_{6,a,b,c}$: the two previous tags and the current tag

SOLUTION

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

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

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

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

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

Exercise 2.

1. Given the training example `the/DT dog/NN saw/VBD the/DT man/NN`, if we convert it to pairs (x, y) pairs, where $x = (y_{i-2}, y_{i-1}, O, i)$ for training a trigram-based CRF model for PoS tagging, which of the following pairs are in the training set?

   (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. List all (x, y) pairs that can be generated from this training set. Assume phantom tags $y_{-2} = y_{-1} =$ `START`.

SOLUTION

1. (a) does not match because position 3 is has $y =$ `VBD` and not `NN`. Similarly, (d) is discarded because position 4 has $y =$ `DT` and not `NN`. Pattern (b) is not matched because previous tags y_{i-2}, y_{i-1} for position 3 are `DT` and `NN` respectively, and not `VBD` and `DT`. The only pair matching the trainig set is (c), since the tag for word 3 is $y =$ `VBD`, and the previous tags y_{i-2}, y_{i-1} are `DT` and `NN` respectively.

2. Pairs generated by this training set are:

$$x = (\texttt{START}, \texttt{START}, \texttt{the dog saw the man}, 1); y = \texttt{DT}$$
$$x = (\texttt{START}, \texttt{DT}, \texttt{the dog saw the man}, 2); y = \texttt{NN}$$
$$x = (\texttt{DT}, \texttt{NN}, \texttt{the dog saw the man}, 3); y = \texttt{VBD}$$
$$x = (\texttt{NN}, \texttt{VBD}, \texttt{the dog saw the man}, 4); y = \texttt{DT}$$
$$x = (\texttt{VBD}, \texttt{DT}, \texttt{the dog saw the man}, 5); y = \texttt{NN}$$

Exercise 3.

We want to approach a PoS tagging task with a bigram-based CRF model that will compute the tag for each word as:

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

We have defined the following feature function types:

- Type 1: Current tag is a:
$$f_{1,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } y_i = a \\ 0 & otherwise \end{cases}$$
- Type 2: Current word is capitalized and current tag is a:
$$f_{2,a}(x_{1:n}, i, y_{i-1}, y_i) = \begin{cases} 1 & \text{if } x_i \text{ is capitalized and } y_i = a \\ 0 & \text{otherwise} \end{cases}$$
- Type 3: Previous tag is a and current tag is 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. Propose values for appropriate features in vector w that will correctly classify all words in the following sentences. Try to set the minimum number of non-zero weights. Proof or justification of the chosen values is required.

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

SOLUTION

- $w_{1,\mathtt{V}} = 1$ will score $+1$ for any word to be tagged as a verb. This will solve correctly all verbs, and introduce a wrong biass 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 begining 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 begining, so we use $+3$.
- $w_{2,\mathtt{E}} = 4$ 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**	E-E-E: 4+0=4	
		N-E: 3+0=3	E-N-E: 4+0=4	
		V-E 1+0=1	**E-V-E: 5+0=5**	
N	N:3	**E-N: 4+0=4**	E-E-N: 4+0=0	
		N-N: 3+0=3	E-N-N: 4+0=0	
		V-N: 1+2=3	**E-V-N: 5+2=7**	←Best
V	V: 1	**E-V: 4+1=5**	E-E-V: 4+1=5	
		N-V: 3+1=4	E-N-V: 4+1=5	
		V-V: 1+1=2	**E-V-V: 5+1=6**	

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**	E-E-E: 4+4=8	
		N-E: 3+0=3	E-N-E: 4+4=8	
		V-E 1+0=1	**E-V-E: 5+4=9**	←Best
N	N: 3	**E-N: 4+0=4**	E-E-N: 4+0=0	
		N-N: 3+0=3	E-N-N: 4+0=0	
		V-N: 1+2=3	**E-V-N: 5+2=7**	
V	V:1	**E-V: 4+1=5**	E-E-V: 4+1=5	
		N-V: 3+1=4	E-N-V: 4+1=5	
		V-V: 1+1=2	**E-V-V: 5+1=6**	

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-E: 3+0=3	
		N-E: 3+0=3	N-N-E: 3+0=3	
		V-E 1+0=1	**N-V-E: 4+0=4**	
N	N: 3	E-N: 0+0=0	N-E-N: 3+0=3	
		N-N: 3+0=3	N-N-N: 3+0=0	
		V-N: 1+2=3	**N-V-N: 4+2=6**	←Best
V	V: 1	E-V: 0+1=1	N-E-V: 3+1=4	
		N-V: 3+1=4	N-N-V: 3+1=4	
		V-V: 1+1=2	**N-V-V: 4+1=5**	

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

Exercise 4.

We are performing PoS tagging with a trigram-factored CRF, using tagset $\mathcal{T} = \{\texttt{DT}, \texttt{V}, \texttt{NN}, \texttt{ADV}, \texttt{PREP}\}$, and we defined a history as $h = \langle t_{i-2}, t_{i-1}, w_{[1:n]}, i \rangle$.

1. How many possible histories are there for a given input sequence $\mathcal{X}$ and a fixed value of i?
2. Which of the following are valid features?

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

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

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

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

3. Compute the global feature vector $\mathbf{f}(\mathcal{X}, \mathcal{Y})$ for the input sequence is $\mathcal{X} =$`the dog walked to a park` and the tag sequence $\mathcal{Y} =$ `DT NN V PREP DT NN`, when using the following features:

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

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

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

4. Given the history $h = (t_{i-2}, t_{i-1}, w_{[1:n]}, 5) = (\texttt{V}, \texttt{DT}, \texttt{the man saw the dog in the park}, 5)$, which of the following features yield $\mathbf{f}(h, \texttt{NN}) = 1$?

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

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

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

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

SOLUTION

1. Each history has the form $h = \langle t_{i-2}, t_{i-1}, w_{[1:n]}, i \rangle$. If we fix $\mathcal{X} = w_{[1:n]}$ and the position i, there are only two parameters left: t_{i-2} and t_{i-1}. Since each of them can take any of the possible five PoS tag values $\{\texttt{DT}, \texttt{V}, \texttt{NN}, \texttt{ADV}, \texttt{PREP}\}$, the number of posible combinations is $5 \times 5 = 25$.
2. Features f_1 and f_2 are valid beacuse they use elements in h (i.e. t_{i-1} and w_{i-2}, respectively). Feature f_3 is invalid because t_{i-3} is not included in h. Feature f_4 is not valid beacuse t_{i+1} is not included in h
3. Given the values of $\mathcal{X}$ and $\mathcal{Y}$, for each word we obtain the following features:

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

Thus, the vector resulting of applying given features is: $(f_1, f_2, f_3) = (1, 2, 1)$

4. Position $i = 5$ corresponds to word `dog` in the sentence. Thus, when evaluating each feature for $t = \texttt{NN}$, we get that:

   $f_1(h, \texttt{NN}) = 1$, since $t = \texttt{NN}$ and $w_5 = \texttt{dog}$
   $f_2(h, \texttt{NN}) = 0$, since $t \neq \texttt{DT}$
   $f_3(h, \texttt{NN}) = 0$, since $w_6 \neq \texttt{dog}$
   $f_4(h, \texttt{NN}) = 1$, since $t = \texttt{NN}$ and $t_{i-1} = \texttt{DT}$

Exercise 5.

We want to address a Named Entity Recognition task consisting in identifying diseases in medical texts. For this, we want to train a sequence classifier such as a CRF using bigram factorization (i.e. only previous and current tag hypothesis are considered). Thus, the used context is $h = (t_{i-1}, w_{[1:n]}, pos_{[1:n]}, i)$.

We use the following feature templates:

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

$$\mathbf{f}_{2}(h,t) = \begin{cases} 1 & \text{if } suf(w_{i-1}) = '\text{ing}' \text{ and } t = \text{B} \\ 0 & \text{otherwise} \end{cases}$$

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

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

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

Given the above templates, and the training sentence:

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

List which feature instances would be generated for words:

a) $i = 2$ (*syndrome*)

b) $i = 10$ (*involving*)

c) $i = 11$ (*mitral*)

SOLUTION

a) Features for $i = 2$ (*syndrome*): $(f_{3,\text{Fragile-X},\texttt{I}}, f_{4,\text{Fragile-X},I,\text{N}}, f_{5,\text{syndrome},I})$

b) Features for $i = 10$ (*involving*): $(f_{1,O}, f_{3,\text{retardation},\texttt{O}}, f_{4,\text{retardation},O,\text{V}})$

c) Features for $i = 11$ (*mitral*): $(f_2, f_{3,\text{involving},\texttt{B}}, f_{4,\text{involving},B,\text{JJ}})$

Exercise 6. Features for log linear sequence annotation models

We are performing PoS tagging for a recently discovered alien language, using a trigram-factored CRF, using tagset $\mathcal{T} = \{\texttt{D}, \texttt{V}, \texttt{N}, \texttt{A}, \texttt{P}\}$, and we defined a history as $h = \langle t_{i-2}, t_{i-1}, w_{[1:n]}, i \rangle$.

1. How many possible histories are there for a given input sequence $\mathcal{X}$ and a fixed value of i? Justify your answer.
2. Which of the following are valid features and which are not? Justify your answer.

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

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

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

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

3. Compute the feature vectors $\mathbf{f}(h,t)$ for each position i, and the global feature vector $\mathbf{f}(\mathcal{X}, \mathcal{Y})$ for the input sequence $\mathcal{X}$ = `grufp umdk wuakla du blha skjkeg` and the tag sequence $\mathcal{Y}$ = `P V N D N A`, when using the following features:

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

$$\mathbf{f}_2(h,t) = \begin{cases} 1 & \text{if } t = \texttt{N} \text{ and } t_{i-1} \neq \texttt{A} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h,t) = \begin{cases} 1 & \text{if } t = \texttt{N} \text{ and } t_{i-1} = \texttt{V} \text{ and } w_{i-1} = \texttt{umdk} \\ 0 & \text{otherwise} \end{cases}$$

SOLUTION

1. For fixed $\mathcal{X}$ and i, the only varible elements of the history are t_{i-2} and t_{i-1}. Since each of them may have any value in $\mathcal{T}$, the number of different possible histories is $|\mathcal{T}|^2 = 5^2 = 25$
2. f_1 is valid, since it depends only on t and t_{i-1}, which is included in h.
   f_2 is not valid, since $\texttt{K} \notin \mathcal{T}$.
   f_3 is not valid, since t_{i-3} is not included in h.
   f_4 is not valid, since t_{i+1} is not included in h.

3. for $i = 1$, we have $h = \langle \texttt{START}, \texttt{START}, \mathcal{X}, 1 \rangle$, and $\mathbf{f}(h,t) = (\mathbf{f}_1(h, \texttt{P}), \mathbf{f}_2(h, \texttt{P}), \mathbf{f}_3(h, \texttt{P})) = (0, 0, 0)$
   for $i = 2$, we have $h = \langle \texttt{START}, \texttt{P}, \mathcal{X}, 2 \rangle$, and $\mathbf{f}(h,t) = (\mathbf{f}_1(h, \texttt{V}), \mathbf{f}_2(h, \texttt{V}), \mathbf{f}_3(h, \texttt{V})) = (0, 0, 0)$
   for $i = 3$, we have $h = \langle \texttt{P}, \texttt{V}, \mathcal{X}, 3 \rangle$, and $\mathbf{f}(h,t) = (\mathbf{f}_1(h, \texttt{N}), \mathbf{f}_2(h, \texttt{N}), \mathbf{f}_3(h, \texttt{N})) = (1, 1, 1)$
   for $i = 4$, we have $h = \langle \texttt{V}, \texttt{N}, \mathcal{X}, 4 \rangle$, and $\mathbf{f}(h,t) = (\mathbf{f}_1(h, \texttt{D}), \mathbf{f}_2(h, \texttt{D}), \mathbf{f}_3(h, \texttt{D})) = (0, 0, 0)$
   for $i = 5$, we have $h = \langle \texttt{N}, \texttt{D}, \mathcal{X}, 5 \rangle$, and $\mathbf{f}(h,t) = (\mathbf{f}_1(h, \texttt{N}), \mathbf{f}_2(h, \texttt{N}), \mathbf{f}_3(h, \texttt{N})) = (0, 1, 0)$
   for $i = 6$, we have $h = \langle \texttt{D}, \texttt{N}, \mathcal{X}, 6 \rangle$, and $\mathbf{f}(h,t) = (\mathbf{f}_1(h, \texttt{A}), \mathbf{f}_2(h, \texttt{A}), \mathbf{f}_3(h, \texttt{A})) = (0, 0, 0)$

   Thus the global feature vector is the sum of the factored vectors: $\mathbf{f}(\mathcal{X}, \mathcal{Y}) = (1, 2, 1)$

Exercise 7.

Negation detection is a task consisting in identifying which phrases in a sentence are affected by a negation. It is a vital task e.g. in applications related to the processing of medical documents.

The task is often modeled as a B-I-O labeling task, and solved using sequence-labeling algorithms such as CRFs.

We have the following training data:

$\mathcal{X}$	The	patient	does	not	show	any	lung	symptoms	.		
$\mathcal{Y}$	O	O	O	B	I	I	I	I	O		
$\mathcal{X}$	Dark	spots	were	observed	in	lung	X-ray	imaging	.		
$\mathcal{Y}$	O	O	O	O	O	O	O	O	O		
$\mathcal{X}$	Exhoglovifin	never	caused	adverse	reactions	and	should	not	be	banned	.
$\mathcal{Y}$	O	B	I	I	I	O	O	B	I	I	O

And the following feature templates:

$$\mathbf{f}_{1,a,l}(\mathcal{X},i,t) = \begin{cases} 1 & \text{if } w_i = a \wedge t = l \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{2,l}(\mathcal{X},i,t) = \begin{cases} 1 & \text{if } w_{i-1} \in \{\textit{no}, \textit{not}, \textit{never}, \textit{any}\} \wedge t = l \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{3,l}(\mathcal{X},i,t) = \begin{cases} 1 & \text{if } \textit{punctuation}(w_i) \wedge\ t = l \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{4,l}(\mathcal{X},i,t) = \begin{cases} 1 & \text{if } w_{i-1} = \textit{dark} \wedge\ w_i = \textit{spots} \wedge\ t = l \\ 0 & \text{otherwise} \end{cases}$$

1. Which is the dimension of the feature space instantiated by this dataset? Justify your answer.
2. Given the following test sentence $\mathcal{X}$ and hypothesis tag sequence $\mathcal{Y}$:

$\mathcal{X}$	X-Ray	results	do	not	show	any	dark	spots	.
$\mathcal{Y}$	O	O	O	B	I	I	I	I	O

compute the feature vectors $\mathbf{f}(\mathcal{X},i,t)$ for each position i, and the global feature vector $\mathbf{f}(\mathcal{X},\mathcal{Y})$. Highlight which features in the global vector that are present in the vector space instantiated by the three training sentences above.

SOLUTION

1. Feature f_1 is instantiated for each combination word-label seen in the training data. Sentence 1 contains 9 combinations. Sentence 2 contains 8 new combinations –combination (.,O) is repeated. Sentence 3 contains 9 new combinations –combinations (not,B) and (.,O) are repeated. Total 9+8+9 = 26 feature instances for template f_1.

   Feature f_2 is instantiated for each occurrence of `not`, `never`, or `any` combined with a label. Sentence 1 contains one occurrence (with label B), sentence 2 does not contain any, and sentence 3 contains two more occurrences, also with label B, so they generate the same feature $f_{2,B}$. Total, 1 feature instances for template f_2.

   Feature f_3 is instantiated for each occurrence of a punctuation sign combined with a label. Each sentence has one occurrence of the combination (.,O), thus only one instance is generated for f_3.

Feature f_4 is instantiated for each occurrence of *dark spots* combined with a label. This only happens once in sentence 2 (with label `O`), thus only one instance is generated for f_4.

So, the total number of generated features (i.e. our feature space dimension) is $26+1+1+1=29$.

2. Feature vectors for each position are:

$$
\begin{aligned}
\mathbf{f}(\mathcal{X},1,\texttt{O}) &= \{f_{1,XRay,O}\}\\
\mathbf{f}(\mathcal{X},2,\texttt{O}) &= \{f_{1,results,O}\}\\
\mathbf{f}(\mathcal{X},3,\texttt{O}) &= \{f_{1,do,O}\}\\
\mathbf{f}(\mathcal{X},4,\texttt{B}) &= \{f_{1,not,B}\}\\
\mathbf{f}(\mathcal{X},5,\texttt{I}) &= \{f_{1,show,I}, f_{2.I}\}\\
\mathbf{f}(\mathcal{X},6,\texttt{I}) &= \{f_{1,any,I}\}\\
\mathbf{f}(\mathcal{X},7,\texttt{I}) &= \{f_{1,dark,I}, f_{2,I}\}\\
\mathbf{f}(\mathcal{X},8,\texttt{I}) &= \{f_{1,spots,I}, f_{4,I}\}\\
\mathbf{f}(\mathcal{X},9,\texttt{O}) &= \{f_{1,.,O}, f_{3,O}\}
\end{aligned}
$$

Thus, the global feature vector $\mathbf{f}(\mathcal{X},\mathcal{Y}) = \sum_i \mathbf{f}(\mathcal{X},i,\mathcal{Y}_i)$ is:

feature	value	in training feature space?
$f_{1,XRay,O}$	1	✓
$f_{1,results,I}$	1	× (word *results* is not in training)
$f_{1,do,O}$	1	× (word *do* is not in training)
$f_{1,not,B}$	1	✓
$f_{1,show,I}$	1	✓
$f_{2,I}$	2	✓
$f_{1,any,I}$	1	✓
$f_{1,dark,I}$	1	× ($f_{1,dark,O}$ appears in training, but $f_{1,dark,I}$ does not)
$f_{1,spots,I}$	1	× ($f_{1,spots,O}$ appears in training, but $f_{1,spots,I}$ does not)
$f_{4,I}$	1	× ($f_{4,O}$ appears in training, but $f_{4,I}$ does not)
$f_{1,.,O}$	1	✓
$f_{3,O}$	1	✓