1 Features in Log-Linear Models (Part 1)

1.1 Question (time: 10:48, slide: 14)

Say we have a set of words $\mathcal{Y}$ with $|\mathcal{Y}| = 1000$, word history $x = \langle w_1 \dots w_{i-1} \rangle$, and let the size of our alphabet be 26 letters.

Define the following features for each word u and suffix s of length 4: $f_{N(u,s)}(x, y) = \begin{cases} 1 & \text{if } y = u \text{ and } w_{i-1} \text{ ends with } s \\ 0 & \text{otherwise} \end{cases}$ where N maps word/suffix pairs to integers.

How many features are there in this model?

(a) 1000×26^4

(b) 26^4

(c) 1000

(d) 1000×26

1.2 Question (time: 13:56, slide: 15)

Say we have a set of words $\mathcal{Y}$ with $|\mathcal{Y}| = 1000$, word history $x = \langle w_1 \dots w_{i-1} \rangle$, and let the size of our alphabet be 26 letters.

Define the following features for each word u and suffix s of length 4: $f_{N(u,s)}(x, y) = \begin{cases} 1 & \text{if } y = u \text{ and } w_{i-1} \text{ ends with } s \\ 0 & \text{otherwise} \end{cases}$ where N maps word/suffix pairs to integers.

Now consider a training set of size $n = 10000$. What is a good upper bound on the number of features introduced for this training set?

(a) 10000

(b) 10000×26^4

(c) 1000×26^4

(d) 1000

2 Features in Log-Linear Models (Part 2)

2.1 Question (time: 7:45, slide: 17)

Consider the label set $\mathcal{Y}$ consisting of part-of-speech tags with $|\mathcal{Y}| = 50$, and the set $\mathcal{X}$ consisting of histories of the form $\langle t_1, \ldots t_i, w_1 \ldots w_n, i \rangle$. Also let the number of words in the vocabulary be 1000.

Say that our features are of the form

$$f_{N(u,t)}(x, y) = \begin{cases} 1 & \text{if } w_i = u \text{ and } y = t \\ 0 & \text{otherwise} \end{cases}$$ where N maps a word/tag pair to an integer.

How many possible features are there in this model?

2.2 Question (time: 7:46, slide: 17)

Consider the label set $\mathcal{Y}$ consisting of part-of-speech tags with $|\mathcal{Y}| = 50$, and the set $\mathcal{X}$ consisting of histories of the form $\langle t_1, \ldots t_i, w_1 \ldots w_n, i \rangle$. Also let the number of words in the vocabulary be 1000.

Say that our features are of the form

$$f_{N(u,t)}(x, y) = \begin{cases} 1 & \text{if } w_i = u \text{ and } y = t \\ 0 & \text{otherwise} \end{cases}$$ where N maps a word/tag pair to an integer.

Our training set consists of the following word/tag pairs

- the/DT man/NN fishes/V for/ADP the/DT fishes/NN

How many different features are introduced in this training set?

3 Definition of Log-linear Models (Part 1)

3.1 Question (time: 4:49, slide: 21)

Consider the label set $\mathcal{Y} = \{\texttt{cat}, \texttt{dog}, \texttt{hat}\}$ with three simple features

- $f_1(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ ends with } \texttt{at} \\ 0 & \text{otherwise} \end{cases}$
- $f_2(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ starts with } \texttt{c} \\ 0 & \text{otherwise} \end{cases}$
- $f_3(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ has second letter } \texttt{o} \\ 0 & \text{otherwise} \end{cases}$

Say we are given the weight vector $v = \langle 1, 2, 3 \rangle$. What is the score of the (x, y) pair $(\texttt{the}, \texttt{cat})$?

4 Definition of Log-linear Models (Part 2)

4.1 Question (time: 0:54, slide: 22)

Consider the label set $\mathcal{Y} = \{\texttt{cat}, \texttt{dog}, \texttt{hat}, \texttt{cot}\}$ with three simple features

- $f_1(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ ends with } \texttt{at} \\ 0 & \text{otherwise} \end{cases}$
- $f_2(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ starts with } \texttt{c} \\ 0 & \text{otherwise} \end{cases}$
- $f_3(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ has second letter } \texttt{o} \\ 0 & \text{otherwise} \end{cases}$

Say we are given the weight vector $v = \langle 0, 0, 0 \rangle$. What is the value of $p(\texttt{cat}|\texttt{the}; v)$?

4.2 Question (time: 0:54, slide: 22)

Consider the label set $\mathcal{Y} = \{\texttt{cat}, \texttt{dog}, \texttt{hat}, \texttt{cot}\}$ with three simple features

- $f_1(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ ends with } \texttt{at} \\ 0 & \text{otherwise} \end{cases}$
- $f_2(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ starts with } \texttt{c} \\ 0 & \text{otherwise} \end{cases}$
- $f_3(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ has second letter } \texttt{o} \\ 0 & \text{otherwise} \end{cases}$

Say we are given the weight vector $v = \langle 3, 1, 1 \rangle$. What is the value of $p(\texttt{hat}|\texttt{the}; v)$ to three decimal places?

5 Parameter Estimation in Log-linear Models (Part 1)

5.1 Question (time: 12:44, slide: 26)

Consider the label set $\mathcal{Y} = \{\texttt{cat}, \texttt{dog}\}$ with three simple features

- $f_1(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ ends with } \texttt{at} \\ 0 & \text{otherwise} \end{cases}$
- $f_2(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ starts with } \texttt{c} \\ 0 & \text{otherwise} \end{cases}$
- $f_3(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ has second letter } \texttt{o} \\ 0 & \text{otherwise} \end{cases}$

Say we are also given two training examples $(x^{(1)}, y^{(1)}) = (\texttt{the}, \texttt{cat})$ and $(x^{(2)}, y^{(2)}) = (\texttt{the}, \texttt{dog})$. If our current weight vector is $v = \langle 0, 0, 0 \rangle$ and $L(v)$ is defined as in the slides, what is the value of $\frac{dL(v)}{dv_2}$?

6 Smoothing Regularization in Log-linear Models

6.1 Question (time: 10:07, slide: 31)

Consider the label set $\mathcal{Y} = \{\texttt{cat}, \texttt{dog}\}$ with three simple features

- $f_1(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ ends with } \texttt{at} \\ 0 & \text{otherwise} \end{cases}$
- $f_2(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ starts with } \texttt{c} \\ 0 & \text{otherwise} \end{cases}$
- $f_3(x, y) = \begin{cases} 1 & \text{if } x = \texttt{the} \text{ and } y \text{ has second letter } \texttt{o} \\ 0 & \text{otherwise} \end{cases}$

Say we are also given two training examples $(x^{(1)}, y^{(1)}) = (\texttt{the}, \texttt{cat})$ and $(x^{(2)}, y^{(2)}) = (\texttt{the}, \texttt{dog})$. If our current weight vector is $v = \langle 1, 1, 1 \rangle$ and $L(v)$ is now defined to use regularization with $\lambda = 1$, what is the value of $\frac{dL(v)}{dv_2}$ to three decimal places?

A Answers

- (a)

  There are 1000 possible values for w that could be in $\mathcal{Y}$ and there are 26^4 possible English suffixes of length four. This gives 1000×26^4 features.

- (a)

  The important new piece of information is the size of the training set. Since there are only 10000 training examples, we will see *at most* 10000 word and suffix pairs, even though 1000×26^4 are possible.

- 50000

  There is one feature for each word and part-of-speech tag. This gives $50 \times 1000 = 50000$ features.

- 5

  There are 5 unique word/tag pairs in the sentence. The pair `the/DT` is used twice. Note that `fishes/V` and `fishes/NN` introduce different features.

- 3

  The first two features are 1 and the third feature is 0. The total score is $\langle 1, 2, 3 \rangle \cdot \langle 1, 1, 0 \rangle = 3$.

- 0.25

  Since the weight vector is 0 all output words have the same score. Therefore all $y \in \mathcal{Y}$, $p(y|\texttt{the}) = 0.25$ and $p(\texttt{cat}|\texttt{the}) = 0.25$

- 0.237

  The score of $(\texttt{the}, \texttt{hat})$ is $v \cdot f(\texttt{the}, \texttt{hat}) = 3$. The sum of exponentiaed scores is $\sum_{y' \in \mathcal{Y}} v \cdot f(\texttt{the}, y') = e^4 + e^1 + e^3 + e^2$. Therefore the conditional probability is $p(\texttt{hat}|\texttt{the}) = e^3/(e^4 + e^1 + e^3 + e^2) = 0.237$.

- 0

  Starting from the definition of the gradient from the slides: $\frac{dL(v)}{dv_2} = \sum_{i=1}^{n} f_2(x^{(i)}, y^{(i)}) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} f_2(x^{(i)}, y') p(y'|x^{(i)}; v) = 1 - (1 \times 0.5 + 0 \times 0.5) - (1 \times 0.5 + 0 \times 0.5) = 0$.

- -1.462

  Note first that $p(\texttt{cat}|\texttt{the}) = e^2/(e^2+e)$ and $p(\texttt{dog}|\texttt{the}) = e/(e^2+e)$. Starting from the definition of the gradient from the slides: $\frac{dL(v)}{dv_2} = \sum_{i=1}^{n} f_2(x^{(i)}, y^{(i)}) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} f_2(x^{(i)}, y') p(y'|x^{(i)}; v) - \lambda v_2 = 1 - \frac{e^2}{e^2+e} - \frac{e^2}{e^2+e} - (1 \times 1) = -\frac{2e^2}{e^2+e} = -1.462$