Conditional Random Fields

Pawan Goyal

CSE, IIT Kharagpur

Week 4, Lecture 5

Practice Question

Suppose you want to use a MaxEnt tagger to tag the sentence, “the light book”. We know that the top 2 POS tags for the words *the*, *light* and *book* are $\{Det, Noun\}$, $\{Verb, Adj\}$ and $\{Verb, Noun\}$, respectively. Assume that the MaxEnt model uses the following history h_i (context) for a word w_i:

$$h_i = \{w_i, w_{i-1}, w_{i+1}, t_{i-1}\}$$

where w_{i-1} and w_{i+1} correspond to the previous and next words and t_{i-1} corresponds to the tag of the previous word. Accordingly, the following features are being used by the MaxEnt model:

- f_1: $t_{i-1} = Det$ and $t_i = Adj$
- f_2: $t_{i-1} = Noun$ and $t_i = Verb$
- f_3: $t_{i-1} = Adj$ and $t_i = Noun$
- f_4: $w_{i-1} = the$ and $t_i = Adj$
- f_5: $w_{i-1} = the \& w_{i+1} = book$ and $t_i = Adj$
- f_6: $w_{i-1} = light$ and $t_i = Noun$
- f_7: $w_{i+1} = light$ and $t_i = Det$
- f_8: $w_{i-1} = NULL$ and $t_i = Noun$

Assume that each feature has a uniform weight of 1.0.
Use Beam search algorithm with a beam-size of 2 to identify the highest probability tag sequence for the sentence.

Problem with Maximum Entropy Models

Per-state normalization

All the mass that arrives at a state must be distributed among the possible successor states

Problem with Maximum Entropy Models

Per-state normalization

All the mass that arrives at a state must be distributed among the possible successor states

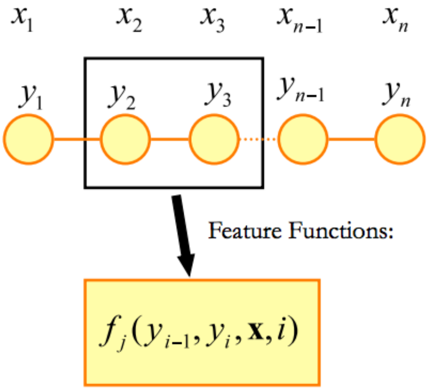
This gives a 'label bias' problem

Let's see the intuition (on paper)

Conditional Random Fields

- CRFs are conditionally trained, undirected graphical models.
- Let's look at the linear chain structure

Conditional Random Fields: Feature Functions



Feature Functions

Express some characteristic of the empirical distribution that we wish to hold in the model distribution

$$f_j(y_{i-1}, y_i, \mathbf{x}, i)$$

1 *if* $y_{i-1} = IN$ *and*
$y_i = NNP$ *and*
$x_i = September$

0 *otherwise*

Conditional Random Fields: Distribution

Label sequence modelled as a normalized product of feature functions:

$$P(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\mathbf{x})} \exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_j (y_{i-1,} y_i, \mathbf{x}, i)$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y} \in Y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_j (y_{i-1,} y_i, \mathbf{x}, i)$$

CRFs

- Have the advantages of MEMM but avoid the label bias problem
- CRFs are globally normalized, whereas MEMMs are locally normalized.
- Widely used and applied. CRFs have been (close to) state-of-the-art in many sequence labeling tasks.