CRF(Conditional random fields)

Prof Nirali Arora

CRF

Conditional Random Fields (CRFs) are a type of statistical modeling method used in Natural Language Processing (NLP) and other fields to handle sequential or structured prediction problems.

They are particularly useful for tasks where you need to predict a sequence of labels or states, such as in part-of-speech tagging, named entity recognition, and other tasks where the output involves labeling each element in a sequence.

CRF : Sequential data handling

Sequential Data Handling: CRFs are designed to model sequential data, where the output is not just a collection of independent labels but has dependencies between neighboring labels. For example, in POS tagging, the label of a word (e.g., noun, verb) depends on the labels of the surrounding words.

Conditional Probability

Conditional Probability: Unlike some other models like Hidden Markov Models (HMMs), which model the joint probability of the observed data and the labels, CRFs model the conditional probability of the labels given the observed data. This conditional approach allows CRFs to better leverage the context of neighboring labels.

Feature Functions

Feature Functions: CRFs use feature functions to capture patterns and relationships in the data. These functions can be designed to capture a wide range of features, including word patterns, neighboring words, and more. The features can be both local (e.g., the current word) and global (e.g., the context of neighboring words).

Graphical model

Graphical Model: CRFs are a type of undirected graphical model. They use a graph structure to represent the dependencies between labels and the relationships between labels and observations. In a CRF, the nodes represent labels or states, and the edges represent dependencies between these labels or states.

Training

Training: Training a CRF involves finding the best set of parameters that maximize the likelihood of the observed data given the labels. This is typically done using optimization techniques like gradient descent and requires algorithms like the forward-backward algorithm and the Viterbi algorithm for efficient computation.

Decoding

Decoding: To make predictions, CRFs use inference techniques to find the most likely sequence of labels given the observed data. The Viterbi algorithm is often used for this purpose, as it efficiently finds the most probable sequence of states.

In summary, CRFs are powerful for tasks involving sequential or structured data because they can model complex dependencies between labels and handle a rich set of features. They are particularly effective in scenarios where the output labels are not independent but have a structured relationship with each other.

Feature functions in CRF

In CRF each feature function is a function that takes in as input
A sentence s
The position i of a word in a sentence

The label li of the current word

The label li-1 of the previous word

And outputs a real valued number

For example ,one possible feature function could be the measure of how much we suspect that the current word can be labelled as adjective given that the previous word is very.

Length of sequence x

Sum over all feature functions

Weight for given feature function

Feature functions

$$P(\mathbf{y}|\mathbf{x};\mathbf{w}) = \frac{\prod_{t=1}^{T} exp\{\sum_{k=1}^{K} w_k \cdot f_k(y_t, y_{t-1}, \mathbf{x})\}}{\sum_{\mathbf{y}} \prod_{t=1}^{T} exp\{\sum_{k=1}^{K} w_k \cdot f_k(y_t, y_{t-1}, \mathbf{x})\}}$$

Sum over all possible label sequences

Feature function can access all observations

$$p(\mathbf{y}|\mathbf{x}) = \underbrace{\frac{1}{Z(\mathbf{x})}}_{\text{Normalization}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \underbrace{\theta_k f_k}_{\text{Weight}} \underbrace{(y_t, y_{t-1}, \mathbf{x}_t)}_{\text{Feature}} \right\}$$

CRF vs HMM

CRF can define much larger set of features

CRF can define both local and global features

HMM is convinced to local features

CRF can have arbitrary weights whereas HMM has to satisfy constraints.