Conditional Random Fields

Shrey Satapara

AI22MTECH02003

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

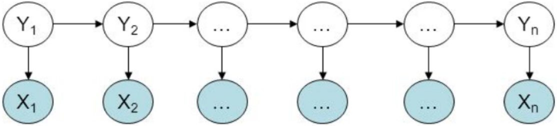
Department of Artificial Intelligence

April 25, 2022

About Referred Paper

- Title: Corpus for Automatic Structuring of Legal Documents Kalamkar et al. 2022
- For segmentation of legal judgment documents in English into topical and coherent parts. Each of these parts is annotated with a label coming from a list of pre-defined Rhetorical Roles. They proposed automatically predicting rhetorical roles using SciBERT-HSLN architecture. Which use CRF as final classification layer. Brack et al. 2021.
- So here, I've explained Conditional Random Fields.

Motivation for CRF: Shortcomings of Hidden Markov Model



- HMM Models direct dependence between each state and only it's corresponding observation.
  - NLP Example: In sequence sentence classification task, target may not only depend on current sentence but also target label of previous sentence or previous sentence.
- Mismatch between learning objective function and prediction objective function.
  - HMM learns a joint distribution of state and observation $P(X, Y)$ but in a prediction task, we need conditional probability $P(Y|X)$

Solution: Conditional Random Fields(CRF)

- To overcome the problem of HMM, CRF came into picture.
- Conditional Random Fields are a type of Discriminator classifier, and as such, they model the decision boundary between the different classes.
- In CRFs, our input data is sequential, and we have to take previous context into account when making predictions on a data point. To model this behavior, we will use Feature Functions, that will have multiple input values, which are going to be:
    - The set of input vectors, X
    - The position i of the data point we are predicting
    - The label of data point i-1 in X
    - The label of data point i in X
- We define feature function as $f(X, i, l_{i-1}, l_i)$

What is feature function

- The purpose of the feature function is to express some kind of characteristic of the sequence that the data point represents.
- For instance, if we are using CRFs for Parts-of-Speach tagging, than
  - $f(X, i, L_{i-1}, L_i) = 1$ if L_{i-1} is a Noun, and $L_i$is a Verb. 0 otherwise.
  - Similarly, $f(X, i, L_{i-1}, L_i) = 1$ if L_{i-1} is a Verb and L_i is an Adverb. 0 otherwise.

Probability Distribution for Conditional Random Fields

- Each feature function is based on previous label and current input. To build conditional field we assign each feature function some weight(lambda values) which our algorithm is going to learn.

$$P(y, X, \lambda) = \frac{1}{Z(X)} exp \Sigma_{i=1}^{n} \Sigma_j \lambda_j f_i(X, i, y_{i-1}, y_i) \tag{1}$$

Where,

$$Z(X) = \Sigma_{y' \in y} \Sigma_{i=1}^{n} \Sigma_j \lambda_j f_i(X, i, y_{i-1}, y_i) \tag{2}$$

Parameters Estimation of Weights(Lambda)

- To estimate the parameters (lambda), we will use Maximum Liklihood Estimation. To apply the technique, we will first take the Negative Log of the distribution, to make the partial derivative easier to calculate:

$$L(y, X, \lambda) = -log(\Pi_{k=1}^{m} P(y^k | x^k, \lambda))$$
$$= -\Sigma_{k=1}^{m} log[\frac{1}{Z(X_m)} exp\Sigma_{i=1}^{n} \Sigma_j \lambda_j f_i(X, i, y_{i-1}, y_i)]$$

Negative Log Liklihood of the CRF Probability Distribution

Parameters Estimation of Weights(contd.)

- To apply Maximum Liklihood on the Negative Log function, we will take the argmin (because minimizing the negative will yield the maximum). To find the minimum, we can take the partial derivative with respect to lambda, and get:

$$\partial L(X, y, \lambda) = \frac{-1}{m} \Sigma_{k=1}^{m} F_j(y^k, x^k) + \Sigma_{k=1}^{m} p(y|x^k, \lambda) F_j(y, x^k) \tag{3}$$

where,

$$F_j(y, x) = \Sigma_{i=1}^{n} f_i(X, i, y_{i-1}, y_i) \tag{4}$$

Partial Derivative w.r.t. lambda

Learning of Weights using Gradient Decent Algorithm

- We use the Partial Derivative as a step in Gradient Descent. Gradient Descent updates parameter values iteratively, with a small step, until the values converge. Our final Gradient Descent update equation for CRF is:

$$\lambda = \lambda - \alpha[\Sigma_{k=1}^{m} F_j(y^k, x^k) + \Sigma_{k=1}^{m} p(y|x^k, \lambda) F_j(y, x^k)] \quad (5)$$

where, α is learning rate

Applications of CRF

- Given their ability to model sequential data, CRFs are often used in Natural Language Processing, and have many applications in that area.
- One such application we discussed is Parts-of-Speech tagging. Parts of speech of a sentence rely on previous words, and by using feature functions that take advantage of this, we can use CRFs to learn how to distinguish which words of a sentence correspond to which POS
- Another similar application is Named Entity recognition, or extracting Proper nouns from sentences.
- Other applications include parts-recognition in Images and gene prediction.

References

Brack, Arthur et al. (2021). *Cross-Domain Multi-Task Learning for Sequential Sentence Classification in Research Papers*. eprint: arXiv:2102.06008.

Kalamkar, Prathamesh et al. (2022). *Corpus for Automatic Structuring of Legal Documents*. eprint: arXiv:2201.13125.

THANK YOU