CS 7180 Ayush Singh

PGM Paper Report

CRF & M3N

Conditional Random Fields

CRF is a markov network in the form of a tree that allows for efficient inferential learning or in other words it is a MRF that can be used to perform classification by exploring the graphical structure of the given model. The key part of CRF is it tries to discriminatively model based on conditional probability P(Y|X) instead of joint probability P(Y,X), modelling the standard prediction problem P(Y|X=x) thus getting greater accuracy compared to others but at the same time is also limited to only predicting applications without introducing further changes.

Typically, parameters in generative models are trained to maximize the joint likelihood of training samples requiring strict independence assumptions of the parameters enumerated computationally which becomes intractable once the number of parameters increase which is generally the case in real world scenarios. Max-Margin Markov Networks captures the above properties along with some other models but face another problem know as label bias problem because of which states with low-entropy next state distributions tend to ignore incoming observations.

Now, for effective parameter estimation and proving convergence of CRF authors use two algorithms based off iterative scaling algorithm testing them to showing empirically that not only do CRFs overcome label bias problem meanwhile getting better accuracy than HMMs but if given a little help (pre-observed parameter dependence) perform even better (25% error rate reduction). The only downsides to it being the amount of time it takes to converge and its dependence on input observed parameter dependence to achieve great error rate reductions.

Max Margin Markov Network Fields (M3N)

As mentioned above, one of the things holding CRF back is its dependence of the preset conditional dependence on parameters. M3N is a parameter learning framework for learning parameters that maximize margin which plays a key role in application where the expectation is not only a classification but a multi-classification. It addresses shortcomings of Markov Networks like handling high dimensional feature spaces and a theoretical bound for generalization in structured domains.

The author starts by encoding the structure using a log linear model in a Markov Network and then on maximizes the margin increasing confidence based on a per-label loss. The key insight is mimicking comes while exploiting structure in M3 Networks reducing exponential number of labels to polynomial sized computation and goes further in using sequential minimal optimization like the one used SVM to reduce large number of quadratic coefficients usually difficult for current quadratic program solvers. Lastly, addressing the topic of generalization bound which is still not on par with that of SVM with kernels, by defining the margin loss to be per-label and proving it to be the bound for generalization error plus a term the grows inversely with the margin.

The other paper generalizes this further leaving M3N to be a subset of Maximum Entropy Discriminative MN. It picks where M3N left off addressing the lack of a straightforward probabilistic interpretation of learning scheme and its prediction rule combining maximum entropy learning (a dual form of maximum likelihood learning) with maximum margin learning of Markov network for structured I/O which gives more robust prediction due to an averaging prediction-function based on the learned distribution of models, Bayesian-style regularization that can lead to a model that is simultaneous primal and dual sparse, and allowance of hidden variables and semi-supervised learning based on partially labeled data as well.

The paper shows significant improvements in the performance on OCR handwriting recognition and collective hypertext classification rate compared to the multi-class SVMs (51% lower) which is really good. Author ends on discussing greater applicability of M3N on real world structured data and its extension to higher-order Markov Models.

Conclusion

Both are structured prediction problem sharing Markov Network Fields (undirected probability graphical distribution over random variables) which is a more general framework which acts as base for the chain CRF we saw in the papers. Markov Random Fields are generative models which model the joint probability distribution which opens them to a wide array of applications, while CRF is a discriminative model which models the conditional probability distribution offering higher accuracy at the cost of requiring pre-defined parameter correlation inference. Although, The intractability is still an issue in this case and using CRFs in high-dimensional feature space but the improvements made and view provided is radical and elegant in borrowing key elements.