Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms

Michael Collins

AT&T Labs-Research Florham Park, New Jersey.

Email address: mcollins@research.att.com

Review by: Jake Pitkin

INTRODUCTION

In this paper Collins described the approach of using variants of the perceptron algorithm for training tagging models. This was an alternative to maximum-entropy models or conditional random fields (CRFs). Using variants of perceptron for sequence tagging problems was an innovative idea as previous to this paper it was mainly used for classification tasks and had not been used for sequence tagging. He conducted experiments for POS tagging and NP chunking, but showed that this new algorithm generalizes to any sequence tagging task that uses Viterbi-style algorithms for decoding.

STRENGTHS OF THIS APPROACH

Collins conducted experiments on two NLP tasks: POS tagging and NP chunking. For each of these tasks, he compared two variants of perceptron (averaged and classic) and a maximum-entropy model. He showed that the use of an averaged perceptron over the then popular ME model gave accuracy improvements. For POS tagging he saw a 11.9% relative reduction in error and a 5.1% relative reduction in error for NP chunking.

In addition to experimental results, he laid out theoretical results that speak in favor of using perceptron variants. He discussed two problems: dealing with inseparable data and producing a classifier that generalizes well on future sequences. Let $\mathbf{U}$ be the vector that separates the training examples with some margin δ, Collins provided a theorem (Theorem 2 Collins 2002) that proves if *most* of the training examples are separable by $\mathbf{U}$ (with some margin δ), then the classifier will make a small number of mistakes on future examples. He also pointed out that there are pre-existing theoretical results suggesting that if perceptron makes a relatively small number of mistakes during training, it is likely to generalize well to new examples. These results appeared in (Freund & Schapire 99) and (Helmbold & Warmuth 95).

What I found to be the strongest strength of this approach is that the number of mistakes the perceptron algorithm makes is independent of the number of possible tag sequences for an input sequence. Rather it depends entirely on the separability of the training data. I think this is powerful as many NLP tasks such as POS tagging can have dozens of possible tags. This leads to an exponential in the size of the input number of possible tag sequences.

EXPERIMENTS

The experiments Collins ran to compare variants of perceptron and maximum-entropy models I found to be fair. He noticed that ME models performance suffers when including rare features, where perceptron variants do not suffer. To accommodate this, he provided results that use all

features and results where a feature is included only if it appears at least five times across the training examples.

An area I found the experiments were lacking is how he presented his results. For POS tagging he provided error rate % as the only comparison metric. I think it would of been more expressive to include the precision and recall for perhaps the five most commons POS tags found in the training examples. For NP chunking he did provide an overall F-measure, but I think it would of been worthwhile to provide more detailed results (error rate, precision, recall) for each of the NP chunking tags (`B, I, O`).

FURTHER DEVELOPMENT

This paper presented the novel idea of using variants of the perceptron algorithm for NLP tasks such as tagging or parsing. The paper spends most of it's time outlining the purposed algorithms, providing theoretical proofs about inseparable training data and generalization, and experimenting with POS tagging and NP chunking.

With this groundwork set, it would be interesting to explore the other applications of using these algorithms anywhere a Viterbi-style algorithm is used for decoding. The author states in theory these algorithms should generalize to these applications, but it would be worthwhile to show this is true in practice.

Collins discussed how theoretical experiment results in (Lafferty et al. 2001) exposed problems with the parameter estimation method for ME models. The response to these problems was an alternative method: Conditional Marjov Random Fields (CRFs), which showed improvements over ME models. My natural response to this is to conduct experiments that compare the results of using CRFs (rather than ME models) and Collins' purposed perceptron algorithms across a variety of NLP sequence tagging tasks.

CLOSING THOUGHTS

Overall I found this paper to be strong as it presents to my knowledge the first use of the perceptron algorithm for a sequence tagging task. It provided theoretical proofs that show the number of mistakes that will be made are independent of the number of possible tag sequences, which can be exponential in the size of the input. Additionally Collins provided experimental results that showed good results on common tasks such as POS tagging or NP chunking, outperforming previous popular methods.

It would be interesting to see work on the comparison between CRFs and the averaged perceptron algorithm on common NLP tasks. Collins states problems ME models face and how CRFs resolve these problems, but conducts his experiments against ME models. Along with expanding the classifiers tested in the experiments, more metrics would of helped convince me of the practical improvements structured perceptron provides on common NLP tasks.

REFERENCES

Collins, Michael. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms.

Freund, Y. & Schapire, R. (1999). Large Margin Classi- fication using the Perceptron Algorithm. In Machine Learning, 37(3):277–296.

Helmbold, D., and Warmuth, M. On weak learning. Journal of Computer and System Sciences, 50(3):551-573, June 1995.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001.