10702 Project: Risk Analysis for Structured Prediction Algorithms

Wenbo Zhao (wzhao1)
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
wzhao1@andrew.cmu.edu

Abstract

Structured prediction refers to a family of supervised machine learnings tasks involving predicting structured outputs. This project studies four main structured prediction algorithms (1) structured perceptron, (2) structural SVM, (3) Maximum Entropy Markov Models (MEMM), and (4) Conditional Random Fields (CRF). The study analyzes and compares their risk minimization strategies and convergence rates, and discuss the significance of the results and open questions.

1 Introduction

In this project, we study the learning algorithms for structured prediction tasks, i.e., structured prediction algorithms, and analyze their risk minimization strategies and convergence rates in probabilistic framework.

Structured prediction refers to a family of supervised machine learnings tasks involving predicting structured outputs, rather than scalar outputs. Such tasks exist in a wide range of applications like natural language processing, speech processing, computer vision, etc. For instance, the task syntactic parsing outputs a grammar-structured tree with its leaves being the words in the given input sequence[3], the task image segmentation outputs segmented graphs from a given image, the task voice onset time prediction outputs a sequence of voice onset time for a given speech [13].

In order to formalize the structured prediction problem, we first introduce two conditions [3].

Condition 1: In a structured prediction problem, output elements $y \in \mathcal{Y}$ decompose into variable length vectors over a finite set. That is, there is a finite $M \in \mathbb{N}$ such that each $y \in \mathcal{Y}$ can be identified with at least one vector $v_y \in M^{T_y}$, where T_y is the length of the vector.

Condition 2: In a structured prediction problem, the loss function does not decompose over the vectors v_y for $y \in \mathcal{Y}$. In particular, $l(x, y, \hat{y})$ is not invariant under identical permutations of y and $\hat{y}$. Formally, we must make this stronger: there is no vector mapping $y \mapsto v_y$ such that the loss function decomposes, for which $|v_y|$ is polynomial in $|y|$.

The first condition constrains the output form, and the second condition restricts the loss form for structured prediction problems. Now we define the structured prediction problem [3, 5].

Problem Formulation: Consider input objects $\boldsymbol{x} \in \mathcal{X}$ and target labels $\boldsymbol{y} \in \mathcal{Y}$ drawn i.i.d from unknown distribution $p \in \mathcal{P}$. Define the feature functions $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$. Denote the decoder score $\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ with parameter $\boldsymbol{w} \in \mathbb{R}^d$. The decoder predicts $\widehat{\boldsymbol{y}}_{\boldsymbol{w}}$ with the highest score

$$\widehat{\boldsymbol{y}}_{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}). \tag{1}$$

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Define the cost function $l(\boldsymbol{y}, \widehat{\boldsymbol{y}}_{\boldsymbol{w}})$. The objective is to minimize the risk

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p}[l(\boldsymbol{y}, \widehat{\boldsymbol{y}}_{\boldsymbol{w}})]. \tag{2}$$

1.1 Motivation

Despite of the potentially wide application of structured prediction, many questions arise. For example, exact inference is often intractable (NP-hard) due to large output search space, or has nonconvex objectives[4]. On the other hand, approximate inference relaxes some constraints but also faces difficulties like reducing model expressivity and misleading standard learning algorithms [7].

The problems existed both in exact and approximate inferences for structured learning motivate us to study the related research, and analyze the consistency and convergence rates for the state-of-art algorithms.

1.2 Review of Structured Prediction Algorithms

Four classes of algorithms have been proposed for structured prediction: (1) structured perceptron, (2) structural SVM, (3) Maximum Entropy Markov Models (MEMM), (4) Conditional Random Fields (CRF). The structured perceptron solves a feasibility problem that is independent of the cost [2]. The structural SVM uses structural hinge loss which is a convex upper bound to the cost [14]. The CRF uses structural log loss [8]. We will analyze the four classes of algorithms in detail.

2 Milestones

- By the second milestone: a full introduction to the structured prediction problem and its related algorithms with precise mathematical description and precise analysis.
- By the final report: a full paper that explains clearly the problem settings, notation and assumptions, key results, proofs, comments on the meaning of the results and open questions, as well as the connection to the concepts learned from class.

2.1 Reading List

The list will necessarily grow as the study goes deeper.

1. Karmon and Keshet [5]
2. Collins [2]
3. Tsochantaridis et al. [14]
4. McAllester [9]
5. Roller [11]
6. Lafferty et al. [8]
7. Smith and Eisner [12]
8. Hazan et al. [4]
9. Chapelle et al. [1]
10. Keshet et al. [6]
11. Ratliff et al. [10]
12. Kulesza et al. [7]

References

[1] Olivier Chapelle, Chuong B Do, Choon H Teo, Quoc V Le, and Alex J Smola. Tighter bounds for structured estimation. In *Advances in neural information processing systems*, pages 281–288, 2009.

[2] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

[3] Harold Charles Daume. *Practical structured learning techniques for natural language processing*. ProQuest, 2006.

[4] Tamir Hazan, Joseph Keshet, and David A McAllester. Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2010.

[5] Danny Karmon and Joseph Keshet. Risk minimization in structured prediction using orbit loss. *arXiv preprint arXiv:1512.02033*, 2015.

[6] Joseph Keshet, David McAllester, and Tamir Hazan. Pac-bayesian approach for minimization of phoneme error rate. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2224–2227. IEEE, 2011.

[7] Alex Kulesza, Fernando Pereira, et al. Structured learning with approximate inference. In *NIPS*, volume 20, pages 785–792, 2007.

[8] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[9] David McAllester. Generalization bounds and consistency for structured labeling. 2009.

[10] Nathan Ratliff, J Andrew Bagnell, and Martin Zinkevich. Subgradient methods for maximum margin structured learning. In *ICML workshop on learning in structured output spaces*, volume 46. Citeseer, 2006.

[11] BTCGD Roller. Max-margin markov networks. *Advances in neural information processing systems*, 16:25, 2004.

[12] David A Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 787–794. Association for Computational Linguistics, 2006.

[13] Morgan Sonderegger and Joseph Keshet. Automatic measurement of voice onset time using discriminative structured prediction a. *The Journal of the Acoustical Society of America*, 132 (6):3965–3979, 2012.

[14] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484, 2005.