ISSN 2186-7437

NII Shonan Meeting Report

No. 2013-6

Discrete Algorithms Meet Machine Learning

Hal Daumé III
Kevin Duh
Samir Khuller

August 10–13, 2013

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Discrete Algorithms Meet Machine Learning

Organizers:
Hal Daumé III (University of Maryland)
Kevin Duh (Nara Institute of Science and Technology)
Samir Khuller (University of Maryland)

August 10–13, 2013

Structured prediction is the task of learning a function that maps inputs to structured outputs, and forms the heart of problems in natural language processing. Syntactic analysis, word alignment for machine translation, semantic understanding, action/goal recognition and translation itself are all examples of structured prediction tasks. In all these tasks, the goal at prediction time is to produce a combinatorial structure, typically by exploiting an off-the-shelf or hand-rolled combinatorial optimization algorithm. The fact that a combinatorial optimization algorithm consumes the output of a machine learning system in order to make predictions renders the learning problem difficult: one must reason about statements like if I change my model in such a way, how will this affect the matching that the Hungarian algorithm (for weighted matchings) would return? Our goal is to solve such questions not only for specific problems such as graphs but also to develop a robust framework in the language of modern combinatorial algorithms by generalizing from spanning trees to matroids, or matchings to matroid intersections. All of these frameworks are polynomial-time for making predictions, and our goal is to construct similar polynomial-time learning procedures.

Although studying polynomial time solvable problems are useful for some specific natural language processing problems, we also wish to turn our attention to approximation algorithms for NP-complete prediction tasks (such as natural language generation or machine translation). So far, such problems have been the bane of structured prediction, requiring either the theoretically ungrounded use of approximate prediction within learning, or efficient but lower-quality search-based solutions with heavy pruning. The key idea we wish to pursue is that NP-complete problems are characterized by efficient verification. We propose to push this further to efficient separation at learning time (which appears to work for all polynomial-time algorithms we have looked at thus far). If this is possible, then we can explicitly train models to give correct solutions, even when a polynomial time approximation algorithm is run at test time. Such a result has the potential to change how the natural language processing community thinks about learning in computationally hard problems.

From the algorithmic perspective, for decades, our basic assumption has been that the input data is sacrosanct and essentially all basic optimization algorithms make this assumption. The problems we envision solving, now question this basic assumption, since we have to modify the input data, while minimizing

the norm of the perturbation, so as to satisfy certain properties. We hope this meeting will unite the large subcommunity in natural language processing community that deals with combinatorial prediction problems with the algorithms community that studies them.

Participants

Hal Daumé III (University of Maryland)
Kevin Duh (Nara Institute of Science and Technology)
Sudipto Guha (University of Pennsylvania)
Satoru Iwata (University of Tokyo)
Samir Khuller (University of Maryland)
David McAllester (Toyota Technological Institute at Chicago)
Julian Mestre (University of Sydney)
Yusuke Miyao (National Institute of Informatics)
Graham Neubig (Nara Institute of Science and Technology)
Emily Pitler (University of Pennsylvania)
Sujith Ravi (Google)
Jun Suzuki (NTT)
Yoshimasa Tsuruoka (University of Tokyo)
Rich Zemel (University Toronto)

Overview of Talks

Recent Advances in Structured Prediction

Daumé III, Hal (University of Maryland)

Machine learning is all about making predictions; structured prediction is about making predictions that have complex, rich internal structure. This tutorial-style talk is all about the how and the why of structured prediction. I will cover the basics of structured prediction: the structured perceptron and incremental parsing and then will then build up to more advanced algorithms that are shockingly reminiscent of these simple approaches: maximum margin techniques and search-based structured prediction.

Interactions between Search and Learning in Statistical Machine Translation

Duh, Kevin (Nara Institute of Science and Technology)

The search problem in Machine Translation is NP-Hard; the learning problem in Machine Translation is full of local optima. This means that I can wait forever for my solution and I don't even know whether it is good or not. In this talk, I will discuss the background about this perennial headache of mine, and explain how our ad hoc techniques deal with it (or not). In particular, I will experimentally demonstrate the effects of hypotheses space approximation on learning and examine the tradeoffs due to N-best enumeration and beam search.

Approximation Algorithms, Indexable Policies and Bayesian Bandit Problems

Guha, Sudipto (University of Pennsylvania)

In this paper, we consider several finite-horizon Bayesian multi-armed bandit problems with side constraints which are computationally intractable (NP-Hard) and for which no optimal (or near optimal) algorithms are known to exist with sub-exponential running time. All of these problems violate the standard exchange property, which assumes that the reward from the play of an arm is not contingent upon when the arm is played. Not only are index policies suboptimal in these contexts, there has been little analysis of such policies in these problem settings. We show that if we consider near-optimal policies, in the sense of approximation algorithms, then there exists (near) index policies. Conceptually, if we can find policies that satisfy an approximate version of the exchange property, namely, that the reward from the play of an arm depends on when the arm is played to within a constant factor, then we have an avenue towards solving these problems. However such an approximate version of the idling bandit property does not hold on a per-play basis and are shown to hold in a global sense. Clearly, such a property is not necessarily true of arbitrary single arm policies and finding such single arm policies is nontrivial. We show that by restricting the state spaces of arms we can find single arm policies and that these single arm policies can be combined into global (near) index policies

where the approximate version of the exchange property is true in expectation. The number of different bandit problems that can be addressed by this technique already demonstrate its wide applicability.

Paper at: arXiv > cs > arXiv:1306.3525: `http://arxiv.org/abs/1306.3525`

Submodular Function Minimization and Bisubmodular Function Maximization

Iwata, Satoru (University of Tokyo)

A function f defined on the subsets of a finite set V is submodular if it satisfies

$$f(X) + f(Y) \geq f(X \cap Y) + f(X \cup Y), \qquad \forall X, Y \subseteq V.$$

Submodular functions are discrete analogues of convex functions. Examples include cut capacity functions, matroid rank functions, and entropy functions.

The first polynomial algorithm for submodular function minimization by Grötschel, Lovász, and Schrijver (1981) is based on the ellipsoid method. In the first part of this talk, I review combinatorial polynomial algorithms for submodular function minimization.

In the second part, I present simple greedy approximation algorithms for maximizing bisubmodular functions, extending the double-greedy algorithms for submodular function maximization due to Buchbinder, Feldman, Naor, and Schwartz (2012). Our deterministic algorithm provides an approximate solution that achieves at least one third of the optimal value, whereas the output of our randomized algorithm achieves at least a half of the optimal value at expectation. This is a joint work with Shin-ichi Tanigawa from Kyoto University and Yuichi Yoshida from NII.

Discrete Algorithms and Structured Prediction

Khuller, Samir (University of Maryland)

Algorithm designers typically assume that the input data is correct, and then proceed to find optimal or sub-optimal solutions using this input data. However this assumption of correct data does not always hold in practice, especially in the context of online learning systems where the objective is to learn appropriate weights given some training samples. Such scenarios necessitate the study of inverse optimization problems where one is given an input instance as well as a desired output and the task is to adjust the input data so that the given output is indeed optimal. In this paper, we consider inverse optimization with a margin, i.e., the given output should be better than all other feasible outputs by a desired margin. We consider such inverse optimization problems for maximum weight matroid basis, matroid intersection, perfect matchings, minimum cost maximum flows, and shortest paths and derive the first known results for such problems with a non-zero margin. The effectiveness of these algorithmic approaches to online prediction is also discussed.

A generalization bound for dropouts

McAllester, David (Toyota Technological Institute at Chicago)

We present a tutorial of the current state of the art of PAC-Bayesian generalization bounds plus a new bound for dropout training. Dropout training has been particularly successful for training deep neural networks and PAC-Bayesian theory is particularly well suited to the analysis of this training method.

Recognizing Textual Entailment by Inference over Algebraic Forms of Sets

Miyao, Yusuke (National Institute of Informatics)

This talk introduces a framework of semantic representation we have proposed recently. The framework is based on algebraic forms of sets, which represent semantics of natural language texts by composing set representations of words. Axioms to infer relations among algebraic forms are defined, and entailment relations among texts are computed as relations among algebraic forms. This framework has been applied to recognizing textual entailment, and we obtained promising results on standard data sets.

Learning a Minimal Rule Set over Packed Forests for Machine Translation

Neubig, Graham (Nara Institute of Science and Technology)

Modern statistical machine translation systems are learned from massive amounts of data, leading to extremely large sets of rules that can be used in translation. In this talk I present some preliminary work on reducing the number of rules to be included in the translation model through the use of machine learning techniques.

Models for Improved Tractability and Accuracy in Dependency Parsing

Pitler, Emily (University of Pennsylvania)

In this talk, we show how characterizations of dependency tree structures can be used to improve the tractability and accuracy of parsing. For languages other than English, dependency parsing has often been formulated as either searching over trees without any crossing dependencies (projective trees) or searching over all directed spanning trees. The former sacrifices the ability to produce many natural language structures; the latter is NP-hard in the presence of features with scopes over siblings or grandparents in the tree. We propose alternative formulations of the dependency parsing problem that include the vast majority of structures seen in treebanks for a variety of natural languages, tractably allow richer features, and have efficient exact parsing algorithms.

We define 1-Endpoint-Crossing trees, in which for any edge that is crossed, all other edges that cross that edge share an endpoint. This property covers

95.8% or more of dependency parses across a variety of languages. We introduce a crossing-sensitive factorization that generalizes a commonly used third-order factorization (capable of scoring triples of edges simultaneously).

Exact dynamic programming algorithms find the optimal 1-Endpoint-Crossing tree under either an edge-factored model or this crossing-sensitive third-order model in $O(n^4)$ time, orders of magnitude faster than other mildly non-projective parsing algorithms and identical to the parsing time for projective trees under the third-order model. The implemented parser is significantly more accurate than the third-order projective parser under many experimental settings and significantly less accurate on none.

Summarization Through Submodularity and Dispersion

Ravi, Sujith (Google)

We propose a new optimization framework for summarization by generalizing the submodular framework of (Lin and Bilmes, 2011). In our framework the summarization desideratum is expressed as a sum of a submodular function and a nonsubmodular function, which we call dispersion; the latter uses intersentence dissimilarities in different ways in order to ensure non-redundancy of the summary. We consider three natural dispersion functions and show that a greedy algorithm can obtain an approximately optimal summary in all three cases. We conduct experiments for two different tasks–single/multi-document summarization and summarizing user comments on news articles–and show that the performance of our algorithm outperforms those that rely only on submodularity.

Supervised Model Learning with Feature Grouping based on a Discrete Constraint

Suzuki, Jun (NTT)

In my talk, we discussed about how we can obtain lower complexity models as possible in supervised learning with a very large feature set, i.e., more than 1 million, which we often encounter in structured prediction problems in NLP field. Currently, a typical technique is to utilize L1-regularizer as a regularization term in the objective function of supervised learning. I introduced a recent developed technique called feature grouping, which is possible to offer much compact models. Unfortunately, the existing feature grouping methods, such as fused lasso and OSCAR regularizers do not work well for very large feature sets, at least, in my preliminary experiments. I proposed a simple but novel idea for feature grouping; to integrate a discrete constraint into model learning. I also introduced a tractable algorithm, which can vanish the intractable combinatorial optimization part from the entire learning algorithm with the help of dual decomposition techniques.

Experiments on two well-studied NLP tasks, dependency parsing and NER, demonstrate that our method can provide models with state-of-the-art performance even if the degrees of freedom in solution models are surprisingly small, i.e., 8 or even 2. There may exist theoretically cleverer approaches to feature grouping, but the performance of our method is already close to the upper

bound. The significant benefit of our method enables us to provide compact model representation, which is especially useful in actual use.

Games, Search and Structured Prediction

Tsuruoka, Yoshimasa (University of Tokyo)

We first describe how evaluation functions for Shogi programs can be trained using a comparison training method, where the parameters of the evaluation function are tuned in such a way that the likelihood of the game records of professional players is maximized. We then talk about how the same technique, which we call lookahead of actions, can be applied to various natural language processing problems such as part-of-speech tagging, named entity recognition, and syntactic parsing. Experimental results demonstrate that this approach gives better performance than standard global models such as conditional random fields. The second topic of the talk is coreference resolution, the goal of which is to group noun phrases that refer to the same real-world entity into one cluster. We propose to perform coreference resolution for pronouns by using a discriminative language model that quantifies the level of naturalness of an expression. We also talk about our recent work on semantic relation classification using recursive neural networks, which are augmented with various pieces of syntactic information obtained from the parse tree and a parameter averaging technique. The final topic of the talk is about the possibility of combining easy-first and lookahead techniques and we present some preliminary results on an English part-of-speech tagging task.

Loss-Sensitive Training Objectives for Probabilistic CRFs

Zemel, Richard (University of Toronto)

We consider the problem of training probabilistic conditional random fields (CRFs) in the context of a task where performance is measured using a specific loss function. While maximum likelihood is the most common approach to training CRFs, it ignores the inherent structure of the tasks loss function. We describe alternatives to maximum likelihood which take that loss into account. These include a novel adaptation of a loss upper bound from the structured SVMs literature to the CRF context, as well as a new loss-inspired KL divergence objective which relies on the probabilistic nature of CRFs. These loss-sensitive objectives are compared to maximum likelihood using ranking as a benchmark task. This comparison confirms the importance of incorporating loss information in the probabilistic training of CRFs, with the loss-inspired KL outperforming all other objectives.