Leveraging Existing Resources using Generalized Expectation Criteria

Gregory Druck
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
gdruck@cs.umass.edu

Gideon Mann
Google, Inc.
76 9th Avenue
New York, NY 10011
gideon.mann@gmail.com

Andrew McCallum
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
mccallum@cs.umass.edu

Abstract

It is difficult to apply machine learning to many real-world tasks because there are no existing labeled instances. In one solution to this problem, a human expert provides instance labels that are used in traditional supervised or semi-supervised training. Instead, we want a solution that allows us to leverage existing resources other than complete labeled instances. We propose the use of generalized expectation (GE) criteria [8] to achieve this goal. A GE criterion is a term in a training objective function that assigns a score to values of a model expectation. In this paper, the expectations are model predicted class distributions conditioned on the presence of selected features, and the score function is the Kullback-Leibler divergence from reference distributions that are estimated using existing resources. We apply this method to the problem of named-entity-recognition, leveraging available lexicons. Using no conventionally labeled instances, we learn a sliding-window multinomial logistic regression model that obtains an F1 score of 0.692 on the CoNLL 2003 data. To attain the same accuracy a supervised classifier requires 4,000 labeled instances.

1 Introduction

Generalized expectation (GE) criteria [8] are terms in a training objective function that assign scores to values of a model expectation. GE resembles the method of moments, but allows us to express arbitrary scalar preferences on expectations of arbitrary functions, rather than requiring equality between sample and model moments. We also note three important differences from traditional training objective functions for factor graphs. First, there need not be a one-to-one relationship between GE terms and model factors. For example, GE allows expectations on sets of variables that form a subset of model factors, or on sets of variables larger than model factors. Next, model expectations for different GE terms can be conditioned on different data sets. Finally, the reference expectation (or more generally, score function) can come from any source, including other tasks or human prior knowledge.

GE provides a method for incorporating prior knowledge into model training. We argue that this approach, in which we communicate with the model by specifying preferences on model expectations, is more intuitive and potentially more robust than incorporating prior knowledge with prior distributions over parameters.

In this paper, we leverage known associations between features and classes. The expectations are model predicted class distributions conditioned on the presence of selected features, and the score function is the Kullback-Leibler divergence from reference distributions that are

estimated using existing resources. Combining these GE terms with a prior on parameters encourages the use of co-occurrence patterns in unlabeled data to learn parameters for features for which we lack prior information.

We apply this method to the task of named-entity-recognition (NER), leveraging existing lexicons, for example lists of universities, organizations, and people. Assuming we have a mapping between lexicons and associated labels, we can estimate reference probability distributions of NER label conditioned on the presence of lexicon features. We use these distributions in a GE objective to train a sliding window multinomial logistic regression classifier. The training procedure requires no labeled sequences.

We provide a discussion of related work (including [10, 1, 3, 9, 4, 5]) elsewhere [2, 6, 8].

2 Generalized Expectation Criterion

A generalized expectation (GE) criterion objective function term assigns scores to values of a model expectation [8]. In many cases this score function is some measure of distance between a model expectation and a reference expectation. Specifically, given some distance function $\Delta(\cdot,\cdot)$, a reference expectation $\hat{f}$, an empirical distribution $\tilde{p}$, a function f, and a conditional model distribution p, the objective function is:

$$\Delta(\hat{f}, E_{\tilde{p}(X)}[E_{p(Y|X;\theta)}[f(X,Y)]]).$$

Here, we explore a special case of GE, similar to the approach of Mann and McCallum [6] and Druck, Mann, and McCallum [2]. More specifically, we use GE in conjunction with multinomial logistic regression models; $\Delta(\cdot,\cdot)$ is the KL divergence; $\tilde{p}$ is unlabeled data; and the expectations are distributions over class conditioned on a specific binary feature, $\tilde{p}(y|x_k = 1;\theta)$, where y is a class label and the feature is indexed by k. We define $\tilde{p}(y|x_k = 1;\theta)$ as:

$$\tilde{p}(y|x_k = 1;\theta) = \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} p(y|\mathbf{x};\theta),$$

where $C_k = \{\mathbf{x} : \tilde{p}(\mathbf{x}) > 0, x_k = 1\}$, the set of all unlabeled instances that contain the feature indexed by k. We discuss the estimation of reference distributions $\hat{p}(y|x_k = 1)$ in the next section. A single GE objective function term is then:

$$D_{KL}(\hat{p}(y|x_k = 1)||\tilde{p}(y|x_k = 1;\theta)) = \sum_y \hat{p}(y|x_k = 1) \log \frac{\hat{p}(y|x_k = 1)}{\tilde{p}(y|x_k = 1;\theta)}.$$

We combine multiple GE terms to obtain the complete objective function:

$$\mathcal{O} = -\sum_{k \in K} D_{KL}(\hat{p}(y|x_k = 1)||\tilde{p}(y|x_k = 1;\theta)) - \frac{\sum_j \theta_j^2}{2\sigma^2}, \quad (1)$$

where K is the subset of features with a GE term. Because there are more parameters in the model than corresponding GE terms in the objective function, the optimization problem is under-constrained. Therefore, we expect there will be many optimal parameter settings. The Gaussian prior addresses this problem by preferring parameter settings with many small values over settings with a few large values. This encourages non-zero parameter values for features that do not have corresponding GE terms, but often co-occur with features with corresponding GE terms. The optimization of Equation 1 is discussed elsewhere [6, 2].

3 Leveraging Existing Resources with GE

To train a model with Equation 1, we need reference distributions $\hat{p}(y|x_k = 1)$. In some scenarios, there exist resources other than conventionally labeled instances that can be used to estimate these distributions. In natural language processing tasks like named-entity recognition, for example, it is common to use readily-available lexicons, or word lists, as features. We often hypothesize associations between these lexicons and labels in the target

task independent of any labeled data. For example, we expect a list of cities to be a good indicator of the *location* label.

We suggest a simple estimation method for converting such lexicon-label associations into distributions over labels conditioned on the presence of label features. For each lexicon feature x_k in an unlabeled instance x, it contributes a vote for each of its associated labels. The instance is assigned the label with the most votes. This results in a labeling of the unlabeled data, and we can compute the distribution $\hat{p}(y|x_k = 1)$ directly.

4 Named-Entity Recognition Experiments

	precision	recall	f1
Association Voting	0.697	0.500	0.582
Logistic Regression w/ GE	**0.725**	**0.662**	**0.692**

Table 1: Named-entity recognition results.

LABEL=ORG	**LABEL=LOC**
POS-TAG=NNPS	WORD=europe
WORD=qantas	WORD=london
WORD=nasdaq	WORD=america
WORD=barcelona	WORD=in@-1
WORD=university	WORD=asia
WORD=ford	WORD=africa
WORD=perth	WORD=america@1
WORD=sydney	WORD=uk
WORD=university@1	WORD=paris
WORD=commonwealth	WORD=south
WORD=rugby	WORD=united
WORD=airways@1	WORD=states@1
WORD=league@1	WORD=bank@-2
POS-TAG=)@-1	WORD=[DATE]@1
WORD=)@-1	POS-TAG=POS@1

Table 2: The most predictive non-lexicon features for the *organization* and *location* labels according to the parameters of a multinomial logistic regression classifier trained with GE.

We evaluate GE training with lexicons on the CoNLL 2003 named-entity recognition data. We formulate the NER problem as a sliding-window classification task. That is, each timestep is an instance, and for each timestep we also include features in a window of ± 3 timesteps. In addition to the word tokens themselves, we include features indicating part-of-speech tags, capitalization, and lexicon matches. More detail on the feature representation is available elsewhere [7]. Note that since we are using a classifier that does not model sequential dependencies, the accuracy cannot be directly compared with linear chain CRFs (for which we are also working on applying GE).

We maximize the objective function Equation 1 with reference distributions estimated using the heuristic described in Section 3. In total, we use 31 lexicons gathered from the web. The model expectations are conditioned on 50,000 unlabeled sentences from the CoNLL 2003 unlabeled data. The results are in Table 1. The model trained with GE achieves an F1 score of 0.692 on the test set (testb). To attain the same accuracy supervised training requires 800 labeled instances of each label type (for a total of 4,000 instances).

We compare against a baseline method that uses directly the association voting method described in Section 3, and GE gives a 26% reduction in F1 error. This comparison illustrates that GE leverages the co-occurrence patterns in unlabeled data during training to learn parameters for non-lexicon features. In Table 2 we show some high-weight parameters for non-lexicon features. Inspection of these features shows that we would indeed expect them to be strong indicators of the label.

5 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by DoD contract #HM1582-06-1-2013, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. In *ACL*, 2007.

[2] Gregory Druck, Gideon Mann, and Andrew McCallum. Reducing annotation effort using generalized expectation criteria. Technical Report 2007-62, University of Massachusetts, Amherst, 2007.

[3] Joao Graça, Kuzman Ganchev, and Ben Taskar. Expectation maximization and posterior constraints. In *NIPS*, 2008.

[4] Aria Haghighi and Dan Klein. Prototype-driven learning for sequence models. In *HTL-NAACL*, 2006.

[5] Rong Jin and Yi Liu. A framework for incorporating class priors into discriminative classification. In *PAKDD*, 2005.

[6] Gideon Mann and Andrew McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, 2007.

[7] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *CoNLL*, 2003.

[8] Andrew McCallum, Gideon Mann, and Gregory Druck. Generalized expectation criteria. Technical Report 2007-60, University of Massachusetts, Amherst, 2007.

[9] Robert E. Schapire, Marie Rochery, Mazin Rahim, and Narendra Gupta. Incorporating prior knowledge into boosting. In *ICML*, 2002.

[10] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin - Madison, 2006.