Empirical Analysis of Collective Stability

Bert Huang[*] BERT@CS.UMD.EDU
Ben London[*] BLONDON@CS.UMD.EDU
Ben Taskar[†] TASKAR@CS.WASHINGTON.EDU
Lise Getoor[*] GETOOR@CS.UMD.EDU
[*] University of Maryland, College Park, MD 20742 USA
[†] University of Washington, Seattle, WA 98195 USA

Abstract

When learning structured predictors, collective stability is an important factor for generalization. London et al. (2013) provide the first analysis of this effect, proving that collectively stable hypotheses produce less deviation between empirical risk and true risk, i.e., defect. We test this effect empirically using a collectively stable variant of max-margin Markov networks. Our experiments on webpage classification validate that increasing the collective stability reduces the defect and can thus lead to lower overall test error.

1. Introduction

London et al. (2013) recently showed that *collective stability*, a measure of a structured predictor's resilience to small perturbations in the input, enables tighter generalization guarantees than those previously known for structured prediction. If every hypothesis in the model class exhibits O(1) collective stability, it can be shown that the empirical risk uniformly converges to the true risk—even in situations where the training set consists of a few large, structured examples. In this work, we empirically evaluate an algorithm inspired by this new theory, demonstrating that the analysis has real, tangible effects on learning to predict structured outputs. In particular, we augment the *max-margin Markov network* (M^{3}N) framework (Taskar et al., 2004) with a collectively stable inference objective to demonstrate that greater collective stability leads to better generalization and can improve overall accuracy, even when learning from a single structured example.

Formally, for a class $\mathcal{F}$ of vector-valued functions, we say that $\mathcal{F}$ has *uniform collective stability* β if, for any two inputs $\mathbf{z}, \mathbf{z}'$ that differ only at a single coordinate,

$$\sup_{f \in \mathcal{F}} \|f(\mathbf{z}) - f(\mathbf{z}')\|_1 \leq \beta.$$

As an example, London et al. (2013) show that certain *templated* models with *strongly convex* inference objectives have collective stability $\mathrm{O}(\sqrt{R/\kappa})$, where R is a bound on the norm of the parameters, and κ is the strong-convexity parameter. For such hypotheses, the defect (i.e., the deviation between the empirical and expected error) is bounded by a quantity with growth rate

$$\mathrm{O}\left(\sqrt{\frac{R \ln n}{\kappa m n}}\right),$$

where m is the number of structured examples, and n is the size of each structured example. (Since here we are concerned with the learning task, we omit from the bound some terms determined by the dependency structure of the true data-generating process; given certain weak dependency conditions, these terms amount to a constant multiplier.) While known approaches minimize the defect by maximizing the margin of a hypothesis (i.e., minimizing the norm of the parameters), the form of this bound suggests another complementary strategy; increasing the strong-convexity parameter—hence, the amount of collective stability—may be yet another tool for reducing the defect. In our experiments, we test whether forcing the hypothesis class to have certain collective stability improves generalization.

2. Max-Margin Markov Networks with Convex Inference

To instantiate a structured predictor where we can experiment with adjusting collective stability, we create

ICML workshop on Structured Learning: Inferring Graphs from Structured and Unstructured Inputs (SLG 2013). Copyright 2013 by the author(s).

a variant of the *max-margin Markov network* ($\mathrm{M}^3\mathrm{N}$) framework (Taskar et al., 2004). The $\mathrm{M}^3\mathrm{N}$ learning algorithm estimates the weights for a log-linear representation of a Markov random field (MRF) by finding a large-margin setting. For observed variables $\mathbf{x}$ and label variables $\mathbf{y}$, a feature-map function $f(\mathbf{x}, \mathbf{y})$ encodes the relevant dependencies, and a weight vector $\mathbf{w}$ defines the conditional log-probability

$$\log \Pr(\mathbf{y} \mid \mathbf{x}; \mathbf{w}) \triangleq \mathbf{w}^\top f(\mathbf{x}, \mathbf{y}) - \log Z.$$

These models are often *templated*, meaning the same weights are applied to all features matching a given dependency pattern. Templating effectively prevents the number of parameters from growing with the size of the input.

Max-margin structured learning aims to find a weight vector $\mathbf{w}$ that puts high probability mass on the ground-truth labels and low probability mass on all other states. In other words, max-margin learning prefers that *maximum a posteriori* (MAP) inference produces accurate predictions. For training with one structured example, the objective function for such a goal is

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \tfrac{1}{2}||\mathbf{w}||^2 + C\xi \\ \text{s.t.} \quad & \mathbf{w}^\top \left(f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \hat{\mathbf{y}})\right) \leq \xi - \ell(\mathbf{y}, \hat{\mathbf{y}}), \forall \mathbf{y} \in \mathcal{Y}, \end{aligned}$$

where ℓ is a loss function, and $\mathcal{Y}$ is the label space (Taskar et al., 2004).

We relax the output space to be a continuous, convex set $\mathcal{A}$, which could be, for example, the marginal or local marginal polytope. We further modify the $\mathrm{M}^3\mathrm{N}$ framework by augmenting the inference objective with a strongly-convex regularization term (i.e., prior). By relaxing the output space to the continuous domain and making inference strongly convex, we guarantee collective stability (London et al., 2013). Though the theory provides guarantees for strong convexity with respect to the 1-norm, in these preliminary experiments, we use a scaled, squared ℓ_2 norm as this strongly-convex term for computational convenience. The inference objective is

$$\max_{\mathbf{y} \in \mathcal{A}} \mathbf{w}^\top f(\mathbf{x}, \mathbf{y}) - \kappa ||\mathbf{y}||^2,$$

where κ is a parameter that adjusts the strong convexity. The max-margin learning objective for this augmented inference is

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \tfrac{1}{2}||\mathbf{w}||^2 + C\xi \\ \text{s.t.} \quad & \mathbf{w}^\top \left(f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \hat{\mathbf{y}})\right) - \kappa \left(||\mathbf{y}||^2 - ||\hat{\mathbf{y}}||^2\right) \\ & \qquad \leq \xi - \ell(\mathbf{y}, \hat{\mathbf{y}}), \forall \mathbf{y} \in \mathcal{A}. \end{aligned}$$

Note that, by setting $\kappa = 0$, inference becomes the linear programming (LP) relaxation of MAP inference, and we obtain the original $\mathrm{M}^3\mathrm{N}$ objective.

We implement collectively stable $\mathrm{M}^3\mathrm{N}$ using a constraint-generation strategy, iteratively finding the worst-violated constraint and adding it to a working set of such constraints. To find the worst-violated constraints, we perform loss-augmented inference, which constructs a quadratic program that maximizes the inference objective subject to local marginal consistency constraints. Since the number of constraints is typically much smaller than the dimensionality of the feature vector, we solve the dual form of the main optimization, which is analogous to the standard dual support vector machine.

3. Experiments

We evaluate collectively stable $\mathrm{M}^3\mathrm{N}$ on the classification of webpages from a subset of the WebKB data set, as preprocessed by Sen et al. (2008). The processed data set consists of networks of webpages belonging to the categories: COURSE, FACULTY, PROJECT, STAFF, and STUDENT. The pages are collected from four universities, and each page is annotated with word occurrences and links. This preprocessed version of the WebKB data set is relatively small, containing on average 219 pages and 402 links per school. We model this data with a Markov network consisting of local, per-page potentials between word occurrence and page category, as well as pairwise edge potentials between all class pairs (Taskar et al., 2002). For max-margin learning, we compute the margin loss only on the singleton label variables, placing no penalty on the pairwise variables.

In each experiment, we train on one university network and test the trained model on the remaining three networks. Since we train each model on a single network, this setup truly tests London et al.'s theory of generalization. We try a variety of slack parameters ($C \in \{0.04, 1, 25\}$) and a range of convexity parameters ($\kappa \in [0, 4]$). Recall that, when κ is zero, we have exactly the standard form of max-margin Markov network learning, since the inference objective becomes the LP relaxation of MAP inference.

To evaluate our predictions, we label each webpage using the most likely category as predicted by the learned model, then compute the classification error rates on both training and testing sets. In Figure 1, we plot the average error rates over four folds of cross-validation. Since we train on one relatively small network at a time, changes in κ and C can cause spurious jumps

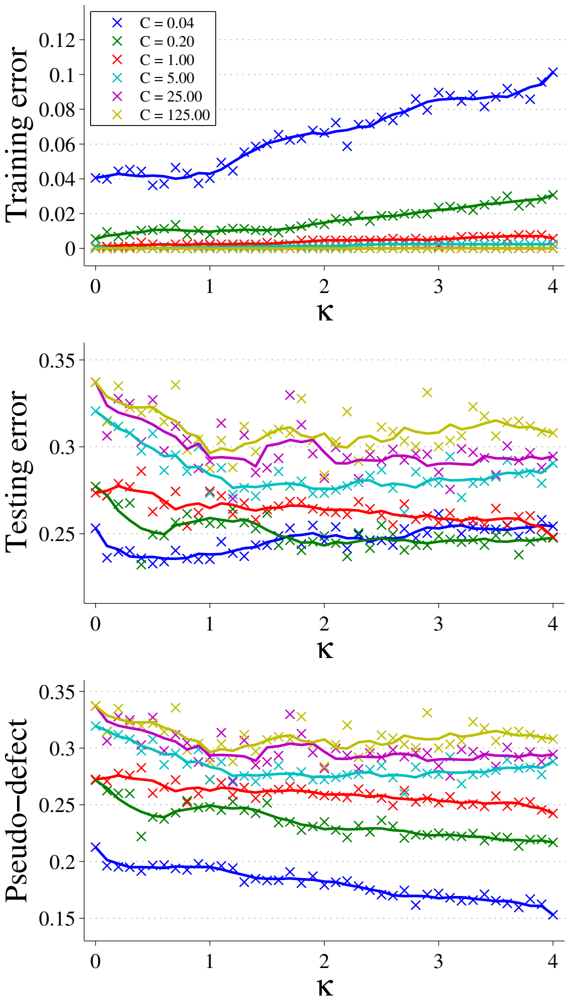

Figure 1. Average training error (top), testing error (middle), and their difference, or pseudo-defect (bottom), as a function of the modulus of convexity, κ. Each point is the average over four folds. The solid lines represent a moving average over a five-point window.

in scores, so we plot a smoothed curve in addition to the point estimates. We compute the smoothed curves by taking the average of a five-point moving window. Examining the accuracies reveals that larger values of κ tend to increase the training error, though when the slack parameter is large, the training error is always near zero. The testing error rates tend to be lower overall with stronger convexity, though the lowest overall testing score is achieved by using a low slack parameter and a small, nonzero convexity. Using no convexity is always worse than using some amount of convexity.

Since the generalization bound analyzes the defect, we measure a surrogate for the defect as the difference of the testing and training errors. We plot the averages of this quantity, which we refer to as *pseudo-defect*, in Figure 1, again including a smoothed version of the curve in addition to the point estimates. The plots show a clear downward trend as the convexity term κ increases.

4. Discussion

In this work, we demonstrate the empirical effect of a recently developed generalization bound for structured prediction (London et al., 2013). Inspired by these bounds, we augment a max-margin structured learning method with a tunable convexity parameter, which effectively controls collective stability of the learned hypothesis. To illustrate the benefits of collective stability, we examine the effect of tuning the convexity parameter during learning. Our experimental results in webpage classification corroborate the importance of collective stability for structured prediction.

Though the theory suggests that an inference objective that is strongly convex with respect to the 1-norm is sufficient for generalization, for the convenience of implementation, our experiments here use an objective that is strongly convex with respect to the 2-norm. The experimental results suggest that even 2-norm strong convexity helps generalization. We are currently investigating approaches to actively optimize the trade-off between collective stability and empirical risk, e.g., by adaptively selecting the strong-convexity term κ. We aim to design an efficiently optimizable learning objective that more closely resembles the risk bound, using entropy-like priors that are strongly convex with respect to the 1-norm. The empirical results presented in this work suggest that finding the optimal κ parameter can provide significant gains in generalization and prediction accuracy.

Acknowledgments

This work was partially supported by NSF CAREER grants 0746930 and 1054215, and NSF grant IIS1218488.

References

London, B., Huang, B., Taskar, B., and Getoor, L. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Taskar, B., Abbeel, P., and Koller, D. Discriminative probabilistic models for relational data. In *Uncertainty in Artificial Intelligence*, 2002.

Taskar, B., Guestrin, C., and Koller, D. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*, 2004.