Probabilistic Soft Logic for Social Good

Stephen H. Bach
Computer Science Dept.
U. of Maryland, College Park
College Park, MD 20742, USA
bach@cs.umd.edu

Bert Huang
Computer Science Dept.
U. of Maryland, College Park
College Park, MD 20742, USA
bert@cs.umd.edu

Lise Getoor
Computer Science Dept.
U. of California, Santa Cruz
Santa Cruz, CA 95064, USA
getoor@soe.ucsc.edu

As governments, non-profit organizations, researchers, and corporations collect data on social phenomena, opportunities have emerged for data science applications that can benefit society. However, modeling these types of complex, real-world phenomena requires new tools to address inherent computational challenges. Social data is intrinsically relational, noisy, partially observed, and large scale, and it is composed of both continuous and discrete information. *Probabilistic soft logic* (PSL) [3, 5] is a general-purpose framework we are developing to solve these challenges.

Since the value of social data is in the networks of relationships they describe, models for social data should be rich enough to capture the intricate dependency structures among unknown variables. These models should be probabilistic so that they are robust to the inconsistencies caused by randomness in the real world while able to capture relevant patterns therein. Further, many variable values may be latent, i.e., not be available for model training, whether because they are inherently unobservable or simply impractical to observe. So, we need methods for learning that support training with latent variables. Finally, because social data is large-scale and contains both discrete and continuous values, we need models that are scalable and support mixtures of discrete and continuous information.

PSL is a declarative language for defining probabilistic models that meet all of these requirements. PSL models are defined as a set of logical rules, each of which describes a common dependency in the data. These rules form templates over entities and relationships in data, enabling interpretable PSL models to reason about complex dependency structures induced by rich, natural networks. For example, in order to detect events in social media, posts and users often have to be accurately *geolocated*, i.e., identified as being posted from or referring to a particular geographic location. A simple PSL rule for geolocation of social-media posts is

$$2.0 \;:\; \textsc{PostMentionsEntity}(P, E) \wedge \textsc{IsLocation}(E) \rightarrow \textsc{HasLocation}(P, E) \;.$$

The rule relates sets of three first-order *atoms*, such as $\textsc{HasLocation}(P, E)$, via probabilistic dependencies for each post P and named entity E. Although the rule certainly does not always hold, PSL will combine it with other rules probabilisitically to make predictions. Each rule is annotated with a non-negative weight, indicating how strongly the rule should hold in the data. To construct a ground model for specific data, each rule is grounded out by replacing the logical variables in the first-order atoms with constants from the data to construct a set of rules containing only ground atoms.

PSL treats the truth values of ground atoms as continuous variables in the $[0, 1]$ interval, using soft logic relaxations of Boolean logic. This continuous representation easily incorporates naturally continuous quantities into its logic-based dependencies, allowing PSL to model and predict both continuous and discrete information by treating the soft truth values as either truly continuous quantities or confidences in discrete predictions. The resulting continuous-variable models are probability densities over the possible continuous truth-value assignments to the ground atoms. Each ground rule induced by a data set contributes a hinge-loss potential function to the graphical model, measuring how far the rule is from being logically satisfied for different assignments of truth values to the ground atoms. These models are members of a powerful class of graphical models known as *hinge-loss Markov random fields* [2], which admit efficient inference and learning in fully-labeled settings as well as partially labeled settings with latent variables.

This intuitive modeling language, backed by scalable machine learning algorithms, makes PSL a flexible tool for social-data applications with the potential for positive impacts. Figure 1 shows an example PSL program for detecting disease outbreaks from social-media posts. The goal is to infer the prevalence of diseases in different locations for a given dictionary of locations and diseases from a corpus of social media posts. The first two rules geolocate posts. The atom $\textsc{PostMentionsEntity}(P, E)$ can be grounded with substitutions via entity recognition on the collection of posts, and $\textsc{IsLocation}(E)$ can be grounded by substituting from the dictionary of locations. $\textsc{PostIsGeotagged}(P, GT)$ can be grounded by extracting any available latitude and longitude geotags from posts, and $\textsc{GeotagInLocation}(GT, L)$ can be grounded via a mapping from geotags to the dictionary of locations. $\textsc{HasLocation}(P, L)$ is unobserved, so the first two rules will be used to infer locations for posts. The locations of posts will then be combined with mentions of diseases in the third rule to infer disease prevalence based on disease mentions. The fourth rule propagates disease prevalence to nearby locations, where $\textsc{Close}(L1, L2)$ is a continuous-valued measure of how geographically close locations are, making this rule propagate disease prevalence more strongly to nearby locations. Finally, the fifth rule acts as prior information, indicating that lower disease prevalence should be preferred in the absence of additional evidence. As additional evidence accumulates, the prior will have less influence on the prediction, and higher prevalence will be predicted.

Previous applications of PSL include detection of events such as disease outbreaks and civil unrest from social me-

$$2.0 \;:\; \textsc{PostMentionsEntity}(P, E) \wedge \textsc{IsLocation}(E) \rightarrow \textsc{HasLocation}(P, E)$$

$$10.0 \;:\; \textsc{PostIsGeotagged}(P, GT) \wedge \textsc{GeotagInLocation}(GT, L) \rightarrow \textsc{HasLocation}(P, L)$$

$$5.0 \;:\; \textsc{PostMentionsDisease}(P, D) \wedge \textsc{HasLocation}(P, L) \rightarrow \textsc{HasDisease}(L, D)$$

$$1.0 \;:\; \textsc{HasDisease}(L1, D) \wedge \text{Close}(L1, L2) \rightarrow \textsc{HasDisease}(L2, D)$$

$$0.5 \;:\; \neg\textsc{HasDisease}(L, D)$$

Figure 1: A sample PSL program for disease-outbreak detection using social media.

dia [6], modeling different types of student engagement in massive open online courses (MOOCs) as latent variables in order to predict outcomes [7], identification of latent group affiliation in social media [1], and predicting trust in social networks [4]. In all of these applications, PSL is applied to social data to make predictions that can have positive impacts. In future work, we are eager to apply PSL to new social-good applications.

We have implemented an open-source software package for the PSL framework[1]. The code is written in Java, so it is portable to a variety of platforms. A Groovy front-end layer is also implemented, allowing users to mix Java and a domain-specific language for easily defining PSL rules and constraints. The PSL package uses a relational-database backend for fast grounding of models. A variety of learning algorithms are included for supervised and semi-supervised learning. We also have implemented an especially scalable algorithm for inference that lazily constructs ground PSL models as it becomes necessary to actually reason about non-zero truth assignments to ground atoms. In relational domains, in which many possible relations do not actually exist, this can greatly improve scalability. The entire package is licensed under the Apache 2.0 license.

In addition to exploring new applications, we are actively researching a number of methodological directions toward making PSL more powerful. We are investigating fast learning algorithms for training from large-scale data that exploit the efficient and convex form of PSL inference. We seek to extend our preliminary work on distributed computation for PSL inference via vertex programming. We are also developing new learning algorithms that are able to learn PSL rules from data, i.e., *structure learning*. Another area we are investigating is using PSL as a decision-support system, helping to target positive interventions by modeling causal relationships. Finally, we are designing methods for learning and inference in streaming settings, in which the model and predictions must be continually—and efficiently—updated as new data arrives.

We have described PSL and some of its qualities that make it well suited to innovative applications of social data with the potential for positive social impacts. PSL's intuitive modeling language and scalable algorithms enable it to work with large-scale social data. With our ongoing research in both algorithms and applications, we are eager to apply PSL to new social-good problems.

[1] `http://psl.cs.umd.edu`

Acknowledgments

This work was supported by NSF grants CCF0937094 and IIS1218488, and IARPA via DoI/NBC contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] S. H. Bach, B. Huang, and L. Getoor. Learning latent groups with hinge-loss Markov random fields. In *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013.

[2] S. H. Bach, B. Huang, B. London, and L. Getoor. Hinge-loss Markov random fields: Convex inference for structured prediction. In *UAI*, 2013.

[3] M. Broecheler, L. Mihalkova, and L. Getoor. Probabilistic similarity logic. In *UAI*, 2010.

[4] B. Huang, A. Kimmig, L. Getoor, and J. Golbeck. A flexible framework for probabilistic models of social trust. In *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction*, 2013.

[5] A. Kimmig, S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

[6] N. Ramakrishnan, P. Butler, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, C. Kuhlman, A. Marathe, L. Zhao, H. Ting, B. Huang, A. Srinivasan, K. Trinh, L. Getoor, G. Katz, A. Doyle, C. Ackermann, I. Zavorin, J. Ford, K. Summers, Y. Fayed, J. Arredondo, D. Gupta, and D. Mares. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *KDD*, 2014.

[7] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *AAAI*, 2014.