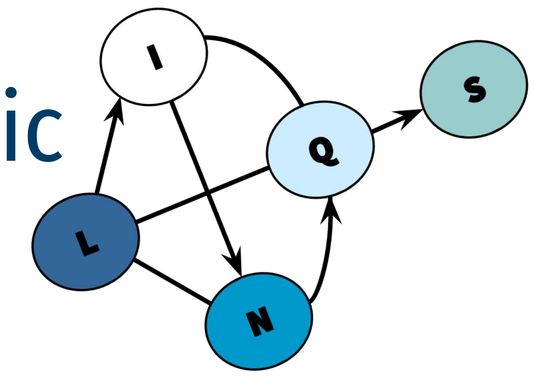
BOWL: Bayesian Optimization For Weight Learning in Probabilistic Soft Logic

Sriram Srinivasan⁺, Golnoosh Farnadi*, & Lise Getoor⁺
⁺ University of California, Santa Cruz
* Mila



Introduction

- Statistical relational learning (SRL) is an effective approach to incorporate relational structure and making collective predictions
- Models often defined through weighted logical rules
- Weights of rules represent importance of the rule
- Weights are important in determining effectiveness of a model
- Previous work focused on finding weights that maximize likelihood

Contributions

- First Bayesian optimization approach for weight learning in SRL
- Show theoretical correctness of our approach
- Empirical evaluation on realworld datasets show that our approach outperforms likelihood based approaches

Background

Probabilistic Soft Logic

- A probabilistic programming language when grounded with data generates a hinge-loss Markov random field (HL-MRF) represented as:

$$P(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{y})} exp(-E(\mathbf{y}|\mathbf{x}))$$

Unobserved random variables Range [0,1] | Observed random variables Range [0,1] | Weights Range $\mathbb{R}^+$ | Normalization constant

Energy function: $\mathbf{E}(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{l} w_i \phi_i(\mathbf{x},\mathbf{y})$ — Number of groundings; Potential function (hinge or linear)

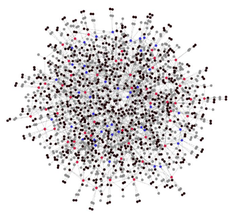
$s.t.,\ \mathbf{0} \le \mathbf{y} \le \mathbf{1}\ ;\ \mathbf{0} \le \mathbf{x} \le \mathbf{1}$

- Inference task translates to convex optimization: $\underset{\mathbf{y}}{\operatorname{argmax}}\, P(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmin}}\, \mathbf{E}(\mathbf{y}|\mathbf{x})$

Example.: simple model for classification task in social network

Weights to be learned: w_1: $LocalPredictor(U) \rightarrow Label(U)$; w_2: $Label(U) \wedge Friend(U, V) \rightarrow Label(U)$

Instantiated with 100 users → Generate HL-MRF



- Existing weight learning application for PSL:
  - MLE: Maximizing likelihood using gradient-based approach
  - MPLE: Maximizing pseudolikelihood using gradient-based approach
  - LME: Large-margin estimation using a loss function and cutting-plane SVM

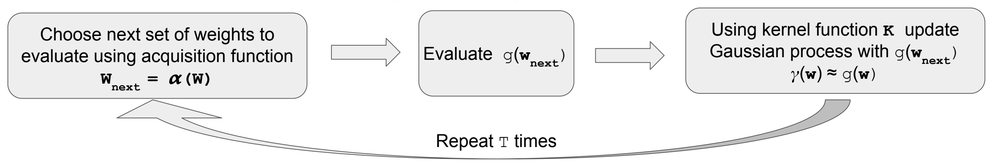
Bayesian Optimization using Gaussian Process Regression

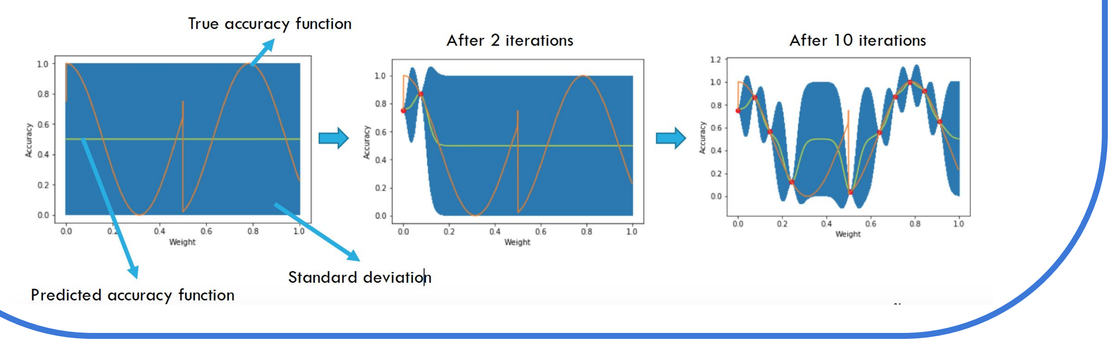
- A sequential design strategy for global optimization of any arbitrary function
- Goal is to find set of weights that maximize any user-defined metric function

$$w^* = argmax_w\ g(w)$$

- Use Gaussian process (GP) regression to approximate the metric function

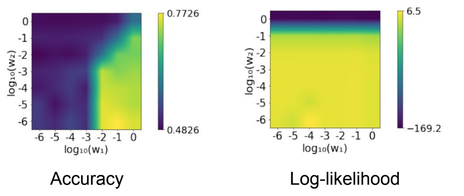
Bayesian optimization:





Motivation For BOWL

- Most approaches maximize likelihood
- We need a generic approach to maximize a user-defined evaluation metric
- Bayesian optimization with GP serves as general framework for such optimization
- How do we effectively optimize for evaluation metric in PSL?



Accuracy | Log-likelihood

Using model defined earlier, we see accuracy is not maximized where log-likelihood is maximized.

BOWL

Original Space (OS)

- Weights re-scaled to [0,1] by dividing $max(w)$
- Squared exponential kernel function:

$$K(w_i, w_j) = \sigma exp\{\delta_{ij}/2\rho^2\}$$

Amplitude | Distance | Length scaling factor

$$\delta_{ij} = ||w_i - w_j||^2$$

BOWL-OS main issue

- Distance does not translate to true correlation between variables
- E.g. : $\boldsymbol{w}_1=\{0.1, 0.1\}$, $\boldsymbol{w}_2=\{1.0, 1.0\}$, $\boldsymbol{w}_3=\{0.1, 0.0001\}$ → $\delta_{12}=1.27$, $\delta_{23}=1.34$, $\delta_{13}=0.09$ — Distance correlation to true distance incorrect

Scaled Space (SS)

- Re-scale weights by projecting weights to a relative weight space
- SS is defined as: $\mathcal{E}(\boldsymbol{w}) \in \mathbb{R}^{(r-1)}$ where r is number of rules

$$\mathcal{E}(\boldsymbol{w}) = \{\forall_{i=2}^{r}(ln(w_i) - ln(w_1))\}$$

Corrected distance: $\Delta_{ij} = ||\mathcal{E}(\boldsymbol{w}_i) - \mathcal{E}(\boldsymbol{w}_j)||^2$

- **Theorem** : **Distance $\Delta_{ij} = 0$ implies that $g(\boldsymbol{w}_i) = g(\boldsymbol{w}_j)$.**
- E.g.: $\Delta_{12}=0.0$, $\Delta_{23}=47.7$, $\Delta_{13}=47.7$: matches our understanding of weights

Effect of grounding in SS

- Grounding adjustment factor for rule z : $\kappa_z = \beta_z / max(\boldsymbol{\beta})$, β_z = # of grounding
- **Theorem:** $||\mathcal{E}(\boldsymbol{w}_i) - \mathcal{E}(\boldsymbol{w}_j)||^2 = ||\mathcal{E}(\boldsymbol{\kappa}.\boldsymbol{w}_i) - \mathcal{E}(\boldsymbol{\kappa}.\boldsymbol{w}_j)||^2$

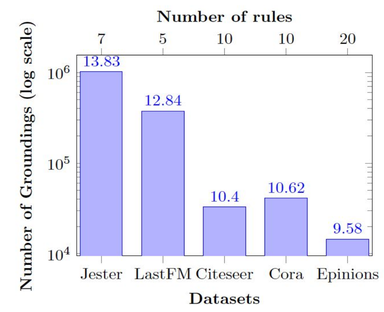
Acquisition functions (α)

- We try four existing acquisition function in Bayesian optimization literature
  - Upper confidence bound (UCB) [Srinivas et al. 2010]
  - Thompson sampling (TS) [Thompson 1933]
  - Probability of improvement (PI) [Kushner 1964]
  - Expected improvement (EI) [Mockus, Tiesis, and Zilinskas 1978]

Empirical Evaluation

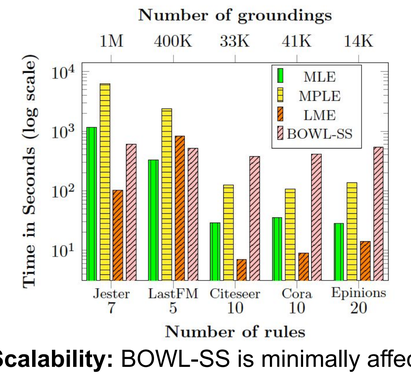
Realworld datasets:

- Jester: Predict user's preference to jokes
  - 2,000 users and 100 jokes.
- LastFM: Predict user's preference to artists
  - 1,892 users and 17,632 artists. The task
- Citeseer: Collective classification on citation dataset
  - 2,708 documents, 5,429 citations, 7 categories.
- Cora: Collective classification on citation dataset
  - 3,312 documents, 4,591 citations, 6 categories.
- Epinions: Predict trust relation between users
  - 2,000 users and 8,675 trust links



Method (Metric)	Jester (MSE)	Jester (AUROC)	LastFM (MSE)	LastFM (AUROC)	Citeeseer (Acc)	Citeeseer (F1)	Cora (Acc)	Cora (F1)	Epinions (AUROC)	Epinions (F1)
MLE	0.058	0.733	0.081	0.581	0.710	0.671	**0.832**	**0.869**	**0.815**	**0.960**
MPLE	0.060	0.737	0.083	0.568	0.729	0.754	**0.832**	**0.869**	0.744	**0.958**
LME	0.055	0.740	0.126	0.554	0.728	0.690	**0.831**	0.849	**0.826**	**0.960**
BOWL-OS	0.055	**0.767**	0.082	**0.599**	**0.740**	**0.796**	**0.832**	**0.869**	**0.812**	**0.962**
BOWL-SS	**0.053**	**0.767**	**0.078**	**0.599**	**0.743**	**0.798**	**0.833**	**0.868**	**0.825**	**0.960**

Performance on user-defined evaluation metric: BOWL-SS significantly outperforms all approaches



Scalability: BOWL-SS is minimally affected by both the number of rules and groundings

Datasets	Different acquisition functions				Varied initializations	
	UCB	TS	PI	EI	Mean	Std
Jester (MSE)	0.053	0.052	0.053	0.053	0.052	0.001
LastFM (MSE)	0.078	0.078	0.078	0.078	0.079	0.001
Citeseer (F1)	0.797	0.797	0.797	0.798	0.804	0.001
Cora (F1)	0.868	0.869	0.866	0.869	0.876	0.002
Epinions (F1)	0.962	0.960	0.960	0.960	0.964	0.002

Robustness: BOWL-SS is unaffected by both acquisition function and initialization

References

- Srinivas, N.; Krause, A.; Kakade, S.; and Seeger, M. 2010. Gaussian process optimization in the bandit setting: No regret and experimental design. In ICML
- Thompson, W. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25(3–4):285–294
- Kushner, H. J. 1964. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. JBE 86 (1):97–106
- Mockus, J.; Tiesis, V.; and Zilinskas, A. 1978. The application of Bayesian methods for seeking the extremum. In TGO