Robust RegBayes

Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models

Shike Mei[1], Jun Zhu[2], Xiaojin Zhu[1]

[1]University of Wisconsin-Madison, [2] Tsinghua University

Motivation

Incorporating knowledge can improve the accuracy (Richardson & Domingos, 2006) and the interpretability of models (Andrzejewski et al., 2011). For a example, in topic model,

Use Domain Knowledge

T1+	T2	T3+	T4	T5+	T6	T7+	T8	T9+	T10
resort	n't	*beach	restaurant	pool	*breakfast	but	*room	*room	hotel
free	pay	*location	fruit	good	*food	n't	told	*bed	*room
*price	but	nice	*dinner	holiday	*service	kids	asked	*bathroom	rooms
great	money	street	wine	bar	but	people	desk	shower	*stay
*worth	check	parking	served	entertainment	day	time	front	*door	hotels
island	time	area	morning	day	water	nice	manager	*floor	night
trip	back	good	menu	*food	bar	night	*stay	colorred*stay	booked
beautiful	car	*restaurant	evening	euros	buffet	great	called	bedroom	*floor
*quality	expensive	internet	meal	lovely	drinks	day	call	coffee	city
place	lobfby	great	eggs	evening	lunch	family	back	towels	view

No Domain Knowledge

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
n't	room	room	room	room	room	hotel	hotel	hotel	hotel
poor	n't	n't	n't	n't	hotel	room	room	room	pool
dirty	told	told	hotel	hotel	n't	n't	n't	day	day
bad	asked	hotel	stay	stay	stay	night	breakfast	staff	area
room	hotel	back	front	night	night	stay	staff	area	staff
hotel	back	front	desk	rooms	rooms	rooms	day	breakfast	rooms
worst	manager	desk	back	back	time	breakfast	night	pool	food
back	stay	stay	night	bed	staff	staff	rooms	time	time
small	called	asked	rooms	front	bed	time	time	n't	breakfast
awful	night	manager	door	time	breakfast	day	area	night	good

Examples of First-Order Logic (FOL) Knowledge

seed-rules:

$\forall i (w(i) = \text{“monkey”}) \rightarrow (z(i) = T)$

cannot-link rules:

$\forall i \forall j (w(i) = \text{“monkey”}) \bigwedge (w(j) = \text{“apple”}) \rightarrow z(i) \neq z(j)$

must-link rules:

$\forall i \forall j (w(i) = \text{“monkey”}) \bigwedge (w(j) = \text{“gorilla”}) \rightarrow z(i) = z(j)$

Contribution

Robust RegBayes framework can incorporate **any FOL knowledge** into **any Bayesian** models as **constraints**.

Previous work incorporate knowledge by prior distribution, which is lacking of flexibility.

Technically, it is a convex framework.

RegBayes: Noiseless knowedge

Standard Bayesian Model:

Prior: $p_0(\mathbf{H})$

Likelihood $p(\mathbf{X}|\mathbf{H})$

FOL base: Knowledge Base containing a set of rules R_l associated with golden standard satisfied proportion γ_l.

$$\gamma_l \triangleq \frac{\sum_{g_l \in G_l} \mathbb{I}_1(g_l(\mathbf{X}, \mathbf{H}))}{|G_l|} \quad (1)$$

Posterior distribution:

$$p(\mathbf{H} \mid \mathbf{X}) \propto p_0(\mathbf{H}) p(\mathbf{X} \mid \mathbf{H})$$

Equivalent Variational Form Bayesian

$$\min_{q \in \mathbb{P}} \mathrm{KL}\left(q(\mathbf{H}) \parallel p(\mathbf{H} \mid \mathbf{X})\right)$$

Define $\phi_l(\mathbf{H}, \mathbf{X})$ as the satisfied proportion of variables on all groundings of an instantiation $(\mathbf{H}, \mathbf{X})$.

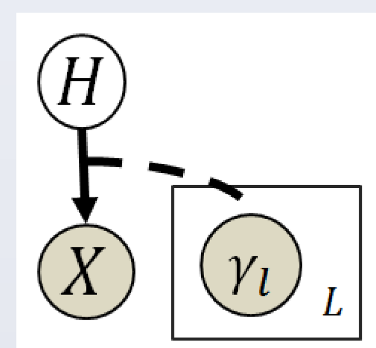
Rules to constrain the variational satisfied proportion $\mathbb{E}_{q(\mathbf{H})}[\phi_l(\mathbf{H}, \mathbf{X})]$ to be close to the golden standard γ_l.

$$|\mathbb{E}_{q(\mathbf{H})}[\phi_l(\mathbf{H}, \mathbf{X})] - \mathbb{E}_{q(\gamma_l)}[\gamma_l]| \leq \epsilon + \xi_l$$

Framework

$$\min_{q, \xi} \quad \mathrm{KL}\left(q(\mathbf{H}) \parallel p(\mathbf{H} \mid \mathbf{X})\right) + C \sum_l \xi_l$$

$$\text{s.t.} \quad |\mathbb{E}_{q(\mathbf{H})}[\phi_l(\mathbf{H}, \mathbf{X})] - \mathbb{E}_{q(\gamma_l)}[\gamma_l]| \leq \epsilon + \xi_l, \quad \xi_l \geq 0, \quad \forall l = 1 \ldots L$$



Optimization

RegBayes is convex! We introduce dual variables $\boldsymbol{\mu}$.

$$q(\mathbf{H} \mid \boldsymbol{\mu}^*) = \frac{p(\mathbf{H} \mid \mathbf{X}, \gamma)}{Z(\boldsymbol{\mu}^*)} e^{\sum_l \mu_l^* (\phi_l(\mathbf{H}, \mathbf{X}) - \gamma_l)}$$

$\boldsymbol{\mu}^*$ is the optimum solution of the dual problem:

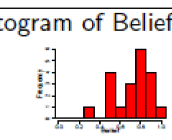
$$\max_{\boldsymbol{\mu}} \quad L(\boldsymbol{\mu}) = -\log Z(\boldsymbol{\mu}) - \epsilon \sum_l \mu_l$$

$$\text{s.t.} \quad -C \leq \mu_l \leq C,$$

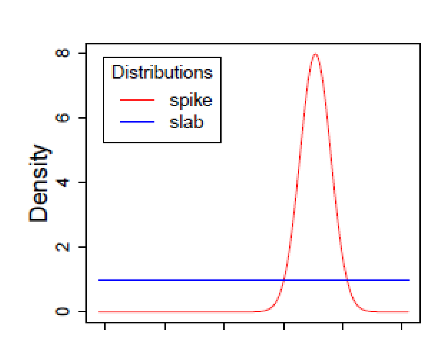
$Z(\boldsymbol{\mu})$ is the normalization factor for q

Robust RegBayes: Noisy Knowledge

Each rule has associated belief labels $\tilde{\gamma}_l$ from M workers: $\tilde{\gamma}_l = \{\tilde{\gamma}_{lm} : \ \tilde{\gamma}_{lm} \in [0, 1]\}_{m=1}^M$.

Histogram of Belief Labels	Rule	Corpus
	$\forall i \forall j (w(i) = human) \bigwedge (w(j) = gene) \rightarrow z(i) \neq z(j)$	Biological Papers

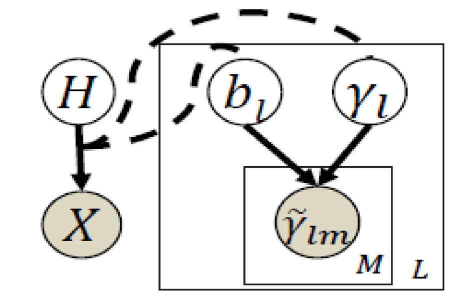
Noisy Belief Likelihood: modeling $p(\tilde{\gamma}_{lm} \mid \gamma_l, b_l)$ as a spike-slab mixture of two components, selected by b_l



Framework

$$\min_{q, \xi} \quad \mathrm{KL}\left(q(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b}) \parallel p(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b} \mid \mathbf{X}, \tilde{\boldsymbol{\gamma}})\right) + C \sum_l \xi_l$$

$$\text{s.t.} \quad \mathbb{E}_{q(b_l)}\left[b_l |\mathbb{E}_{q(\mathbf{H}|b_l)}[\phi_l(\mathbf{H}, \mathbf{X})] - \mathbb{E}_{q(\gamma_l|b_l)}[\gamma_l]|\right] \leq \epsilon + \xi_l, \quad \xi_l \geq 0, \quad \forall l = 1 \ldots L$$



Experiments on Topic Models

Hypothesis one to test

- (Nonrobust) RegBayes can incorporate FOL domain knowledge into topic models.

Hypothesis two to test

- Robust RegBayes can robustly incorporate FOL knowledge.

Hypothesis One : Vanilla LDA

Dataset	#Documents	#Topics	Description	#FOL Rules
COMP	5,000	20	comp.* in 20 newsgroup data	8 seeds
COM	2,740	25	U.S. House of Representatives	3 seeds, 2 docseeds
POL	2,000	20	movie reviews	1 cannot-link
HDG	24,073	50	PubMed abstracts	8 seeds, 6 inclusion, 6 exclusion

Results

Table 2. Test set perplexity and proportion of satisfied logic rules on four datasets.

	Test Set Perplexity			Proportion of Satisfied Logic Rules		
	LDA	Fold·all	LogicLDA	LDA	Fold·all	LogicLDA
COMP	1531 ± 12	1537 ± 11	$\mathbf{1463 \pm 5}$	0.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.01
CON	$\mathbf{1206 \pm 6}$	1535 ± 10	$\mathbf{1216 \pm 11}$	0.07 ± 0.04	0.67 ± 0.03	0.70 ± 0.00
POL	3218 ± 13	3220 ± 13	$\mathbf{3176 \pm 12}$	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
HDG	940 ± 6	973 ± 7	$\mathbf{885 \pm 2}$	0.60 ± 0.01	0.95 ± 0.00	0.96 ± 0.01

Rules' satisfied proportion is high.

Smaller test set perplexity by incorporating domain knowledge (vs LDA and Foldall)

Hypothesis One: Supervised LDA

Task: Given HotelReview dataset, predict the rating of hotel (1 to 5 stars) based on the content of hotel.

Two kinds of domain Knowledge

- Distinguish the topics as related to *value*, *location*, *service* and *room* aspects. Specifically,

Seed words	Topic	Aspect
{*value, price, quality, worth, resort*}	T1-2	*value*
{*location, traffic, restaurant, beach*}	T3	*location*
{*service, food, breakfast, dinner*}	T4-6	*service*
{*door, floor, bed, stay, bathroom, room*}	T7-10	*room*

- Two grammar rules, “Not” rule and “But” rule.

Two Baselines:

(*i*) MedLDAr (Zhu et al.,2013a), a RegBayes model that incorporates max-margin posterior regularization into LDA;

(*ii*) sCTRF (Zhu & Xing, 2010), a feature based model that incorporates both single and pairwise word features into MedLDAr.

Results:

Better Interpretability, see the topic samples in motivation part.

Better Predicting performance.

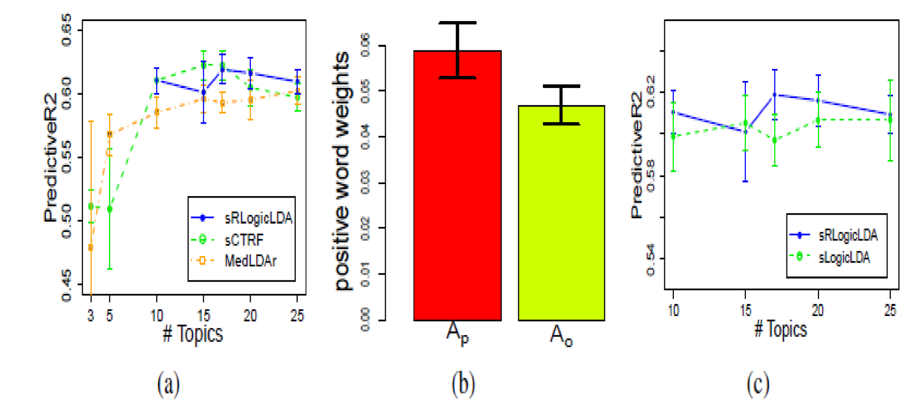
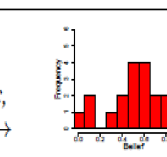
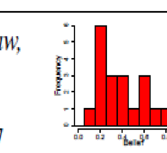
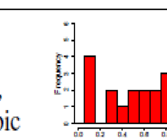
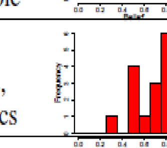


Figure 2. (a) Predictive R^2 of sRLogicLDA, sCTRF, and MedLDAr. (b) average weights of positive words in the positive topic set (A_p) and the other topic set (A_o); and (c) predictive R^2 of sRLogicLDA and sLogicLDA.

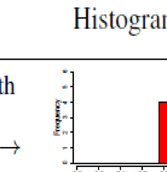
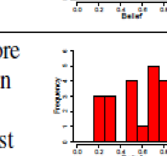
Hypothesis Two: Vanilla LDA

Table 3. RLogicLDA is robust to unreliable domain knowledge

Data	Designed Rule	Histogram	Test Set Perplexity		Satisfaction Proportion	
			LogicLDA	RLogicLDA	LogicLDA	RLogicLDA
COMP	seed: {*problem, windows, window, available, files, mac, apple, system, im*} → topic 1		1467 ± 6	$\mathbf{1446 \pm 6}$	0.39 ± 0.16	0.07 ± 0.06
CON	seed: {*bill, people, law, health, tax, trade, economy, budget, pension*} → topic 7		$\mathbf{1228 \pm 9}$	$\mathbf{1228 \pm 16}$	0.49 ± 0.03	0.08 ± 0.03
POL	must-link{*acting, make, performance, character*}: same topic		$\mathbf{3173 \pm 13}$	$\mathbf{3168 \pm 11}$	0.57 ± 0.19	0.09 ± 0.04
HDG	cannot-link{*human, gene*}: different topics		895 ± 2	$\mathbf{891 \pm 2}$	0.75 ± 0.03	0.95 ± 0.01

Hypothesis Two: Supervised LDA

Table 4. Intentionally Designed Reliable and Unreliable Rules

Rule	Description	Histogram	mean($\tilde{\gamma}_{lm}$)	$p(b_l = 1 \mid \lambda_l)$	Satisfaction Proportion	
					sLogicLDA	sRLogicLDA
Not rule	seed: {adjectives with negation within distance 4 before it} → the last topic		0.91	0.99	0.98 ± 0.03	1.00 ± 0.00
But rule	seed: {all words before adversative transition (e.g. “but”) in sentences} → the last topic		0.56	0.00	0.70 ± 0.13	0.05 ± 0.4

Conclusion

Robust RegBayes framework can robustly incorporate any FOL knowledge into any Bayesian models.

Robust RegBayes improves the model accuracy and interpretability.

ACKNOWLEDGEMENT

The work was supported by the National Basic Research Program of China (Nos. 2013CB329403, 2012CB316301) and National Natural Science Foundation of China (Nos. 61322308, 61332007) to JZ, and a Google Faculty Research Award to XZ.