Robust RegBayes

Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models

Shike Mei[†§], Jun Zhu[§], Xiaojin Zhu[†]

[†]University of Wisconsin-Madison
[§]Tsinghua University

Table of Contents

Motivation

Top words in learned topics from hotel review

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

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**

Motivation

Incorporating knowledge can improve the **accuracy** (Richardson & Domingos, 2006) and the **interpretability** of models (Andrzejewski et al., 2011).

Examples of Knowledge in Topic Modeling

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:

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

Table of Contents

Contributions

Contributions

- Previous work model knowledge as prior, which are lacking of flexibility.

Contributions

- Previous work model knowledge as prior, which are lacking of flexibility.
- Robust RegBayes framework can incorporate **any FOL knowledge** into **any Bayesian models** as soft **constraints**.

Contributions

- Previous work model knowledge as prior, which are lacking of flexibility.
- Robust RegBayes framework can incorporate **any FOL knowledge** into **any Bayesian models** as soft **constraints**.
- Technically, it is a **convex** framework.

Table of Contents

RegBayes

Standard Bayesian Model:
Prior: $p_0(\mathbf{H})$
Likelihood $p(\mathbf{X}|\mathbf{H})$

RegBayes

Standard Bayesian Model:
Prior: $p_0(\mathbf{H})$
Likelihood $p(\mathbf{X}|\mathbf{H})$
(Noiseless) FOL base: Knowledge Base containing a set of rules R_l associated with golden standard satisfied proportion γ_l.

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

E.g. for the seed rule $\forall i(w(i) = \text{"monkey"}) \rightarrow (z(i) = T)$,

$$G_l = \{z(i) = T : w(i) = \text{"monkey"}\}$$

RegBayes

- Posterior distribution:

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

RegBayes

- Posterior distribution:

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

- Equivalent Variational Form Bayesian

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

RegBayes

- Posterior distribution:

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

- Equivalent Variational Form Bayesian

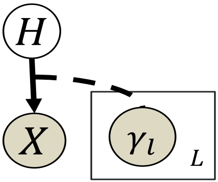
$$\min_{q \in \mathbb{P}} \mathrm{KL}\left(q(\mathbf{H}) \parallel p(\mathbf{H} \mid \mathbf{X})\right)$$

- 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 $\mathbb{E}_{q(\mathbf{H})}\left[\phi_l(\mathbf{H}, \mathbf{X})\right]$ to be close to the golden standard γ_l.

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

RegBayes

$$
\begin{aligned}
\min_{q,\boldsymbol{\xi}} \quad & \mathrm{KL}\left(q(\mathbf{H}) \parallel p(\mathbf{H} \mid \mathbf{X})\right) + C \sum_{l} \xi_l \\
\text{s.t.} \quad & \left|\mathbb{E}_{q(\mathbf{H})}\left[\phi_l(\mathbf{H}, \mathbf{X})\right] - \gamma_l\right| \leq \epsilon + \xi_l, \\
& \xi_l \geq 0, \quad \forall l = 1 \ldots L
\end{aligned}
$$



Optimization

- RegBayes is convex! We introduce dual variables $\boldsymbol{\mu}$ (weights of rules).

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

Optimization

- RegBayes is convex! We introduce dual variables $\boldsymbol{\mu}$ (weights of rules).

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

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

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & L(\boldsymbol{\mu}) = -\log Z(\boldsymbol{\mu}) - \epsilon \sum_l \mu_l \\ \text{s.t.} \quad & -C \leq \mu_l \leq C, \end{aligned}$$

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

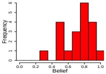
Table of Contents

Obtaining Robustness

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

Obtaining Robustness

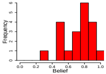
Each rule has associated belief labels $\tilde{\gamma}_l$ from M workers:
$\tilde{\gamma}_l = \{\tilde{\gamma}_{lm} : \ \tilde{\gamma}_{lm} \in [0,1]\}_{m=1}^{M}$.

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

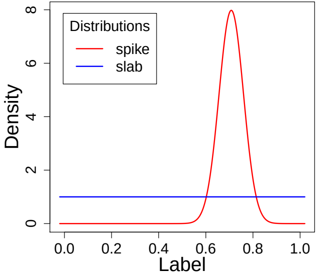
Noisy Belief Likelihood: modeling $p(\tilde{\gamma}_{lm} \mid \gamma_l, b_l)$ as a spike-slab mixture of two components, selected by b_l (reliability)

Obtaining Robustness

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 (reliability)



Obtain Robustness

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

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

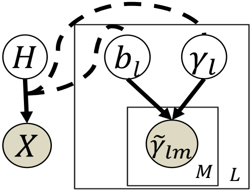


Table of Contents

Experiment

Hypothesis one to test

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

Hypothesis two to test

- Robust RegBayes can **robustly** incorporate FOL knowledge against noise in expert input.

Datasets

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

LogicLDA (RegBayes) vs LDA

	Proportion of Satisfied Logic Rules	
	LDA	LogicLDA
COMP	0.00 ± 0.00	$\mathbf{0.97} \pm 0.01$
CON	0.07 ± 0.04	$\mathbf{0.70} \pm 0.00$
POL	1.00 ± 0.00	$\mathbf{1.00} \pm 0.00$
HDG	0.60 ± 0.01	$\mathbf{0.96} \pm 0.01$
	Test Set Perplexity	
COMP	1531 ± 12	$\mathbf{1463} \pm 5$
CON	$\mathbf{1206} \pm 6$	$\mathbf{1216} \pm 11$
POL	3218 ± 13	$\mathbf{3176} \pm 12$
HDG	940 ± 6	$\mathbf{885} \pm 2$

LogicLDA (RegBayes) vs LDA

	Proportion of Satisfied Logic Rules	
	LDA	LogicLDA
COMP	0.00 ± 0.00	$\mathbf{0.97} \pm 0.01$
CON	0.07 ± 0.04	$\mathbf{0.70} \pm 0.00$
POL	1.00 ± 0.00	$\mathbf{1.00} \pm 0.00$
HDG	0.60 ± 0.01	$\mathbf{0.96} \pm 0.01$
	Test Set Perplexity	
COMP	1531 ± 12	$\mathbf{1463} \pm 5$
CON	$\mathbf{1206} \pm 6$	$\mathbf{1216} \pm 11$
POL	3218 ± 13	$\mathbf{3176} \pm 12$
HDG	940 ± 6	$\mathbf{885} \pm 2$

- Rules' satisfied proportion is high.

LogicLDA (RegBayes) vs LDA

	Proportion of Satisfied Logic Rules	
	LDA	LogicLDA
COMP	0.00 ± 0.00	$\mathbf{0.97} \pm 0.01$
CON	0.07 ± 0.04	$\mathbf{0.70} \pm 0.00$
POL	1.00 ± 0.00	$\mathbf{1.00} \pm 0.00$
HDG	0.60 ± 0.01	$\mathbf{0.96} \pm 0.01$
	Test Set Perplexity	
COMP	1531 ± 12	$\mathbf{1463} \pm 5$
CON	$\mathbf{1206} \pm 6$	$\mathbf{1216} \pm 11$
POL	3218 ± 13	$\mathbf{3176} \pm 12$
HDG	940 ± 6	$\mathbf{885} \pm 2$

- Rules' satisfied proportion is high.
- smaller test set perplexity by incorporating domain knowledge (vs LDA).

Experiment on Supervised LDA

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

Experiment on Supervised LDA

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

Experiment on Supervised LDA

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

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.

Interpretability of Topics

No Domain Knowledge

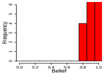
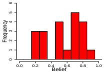
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
n't	room	room	room	room	room	hotel	hotel	hotel	hotel
poor	n't	n't	n't	n't	hotel	room	room	room	pool
dirty	told	told	hotel	hotel	n't	n't	n't	day	day
bad	asked	hotel	stay	stay	stay	night	breakfast	staff	area
room	hotel	back	front	night	night	stay	staff	area	staff
hotel	back	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

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**

sRLogicLDA vs sLogicLDA

sLogicLDA naively constrains the satisfied proportion close to the mean, while sRLogicLDA filters the 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} $\rightarrow$ 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} $\rightarrow$ the last topic		0.56	0.00	0.70 ± 0.13	0.05 ± 0.4

Conclusion

Conclusion

- Robust RegBayes framework can **robustly** incorporate **any FOL knowledge** into any **Bayesian model**.

Conclusion

- Robust RegBayes framework can **robustly** incorporate **any FOL knowledge** into any **Bayesian model**.
- Robust RegBayes framework improves the model accuracy and interpretability.

Conclusion

- Robust RegBayes framework can **robustly** incorporate **any FOL knowledge** into any **Bayesian model**.
- Robust RegBayes framework improves the model accuracy and interpretability.
- More information on poster S14

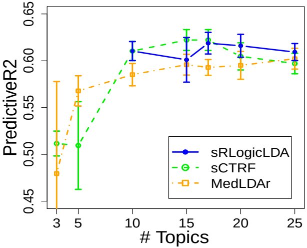
Two Baselines

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

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.

Predictive R^2 of sRLogicLDA, sCTRF, and MedLDAr



- sRLogicLDA achieves the similar performance with sCTRF (with only 7 rules vs 15 features on words in sCTRF) and better predictive performance than MedLDA.
- Incorporating domain knowledge improves predictive accuracy.