MedLDA: Maximum Margin Supervised Topic Models

Jun Zhu

Amr Ahmed

Eric P. Xing

Presented by Zhengming Xing

Outline

- Introduction
- Background
- MedLDA
- Experiment result

Introduction

Main idea:

- Supervised topic modeling (Utilize the side information, such as rating and label information)
- Integrate the mechanism behind max-margin prediction algorithm with hierarchical Bayesian topic modeling.

Advantage:

- Discover sparse and highly discriminative topical representation
- Achieve state of art prediction performance
- More efficient than existing supervised model

Background (Topic modeling)

LDA

1. For a document d, draw a topic mixing proportion vector $\boldsymbol{\theta}_d$ according to a K-dimensional Dirichlet prior: $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha})$;
2. For the n-th word in document d, where $1 \leq n \leq N$,
   (a) draw a topic assignment z_{dn} according to $\boldsymbol{\theta}_d$: $z_{dn}|\boldsymbol{\theta}_d \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$;
   (b) draw the word instance w_{dn} according to z_{dn}: $w_{dn}|z_{dn}, \boldsymbol{\beta} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{dn}})$,

Joint distribution

$$p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \Big(\prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \Big)$$

Supervised LDA

$$p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \Big(\prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \Big) p(y_d|\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d, \delta^2)$$

Side information

$$y|\mathbf{z}_d, \boldsymbol{\eta}, \delta^2 \sim \mathcal{N}(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d, \delta^2) \quad \bar{\mathbf{z}}_d \triangleq \frac{1}{N} \sum_n z_{dn}$$

Background (Topic modeling)

marginal data likelihood

$$P(\mathcal{D}) = \int P(\mathcal{D}, \Omega) d\Omega$$

Variational bound

$$lnP(\mathcal{D}) = \mathcal{L}(Q) + KL(Q||P) \qquad \mathcal{L} = \int Q(\Omega) ln \frac{P(\mathcal{D},\Omega)}{Q(\Omega)} d\Omega$$

LDA

$$\begin{aligned}\mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta}) &\triangleq -\mathbb{E}_q[\log p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})] - \mathcal{H}(q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})) \\ &\geq -\log p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}),\end{aligned}$$

approximate the marginal data likelihood $p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$

sLDA

$$\begin{aligned}\mathcal{L}^s(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2) &\triangleq -\mathbb{E}_q[\log p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)] - \mathcal{H}(q(\{\theta_d, \mathbf{z}_d\})) \\ &\geq -\log p(\mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2).\end{aligned}$$

approximate the marginal data likelihood $p(\mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)$

Background (SVR)

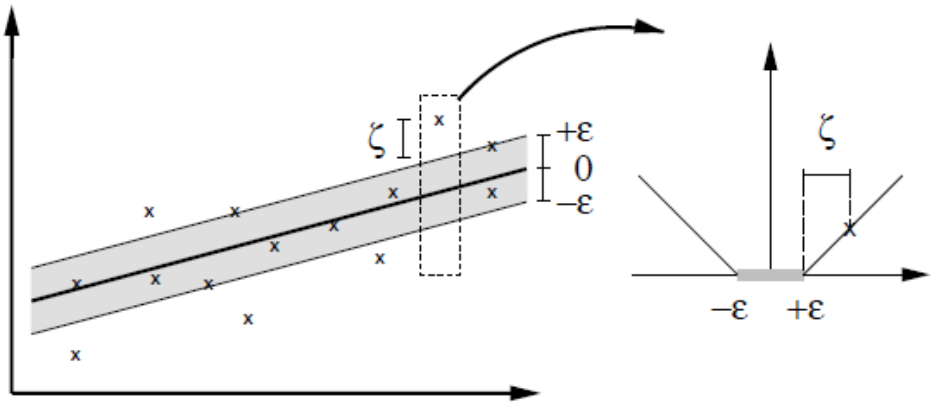
Given the training data

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_D, y_D)\}$$

Objective function

$$\text{P0(SVR)}: \quad \min_{\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\xi}^\star} \quad \frac{1}{2}\|\boldsymbol{\eta}\|_2^2 + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\forall d, \text{ s.t.}: \quad \begin{cases} y_d - \boldsymbol{\eta}^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d \\ -y_d + \boldsymbol{\eta}^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d^\star \\ \xi_d, \xi_d^\star \geq 0 \end{cases},$$



Background (Med)

Maximum entropy discrimination

Find the distribution of all possible regression coefficient belong to a particular distribution class family

Objective function

$$\mathrm{P1(MED^r)}: \quad \min_{q(\boldsymbol{\eta}),\boldsymbol{\xi},\boldsymbol{\xi}^\star} \quad KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\forall d,\ \text{s.t.}: \quad \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d \\ -y_d + \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d^\star \\ \xi_d, \xi_d^\star \geq 0 \end{cases},$$

$p_0(\eta)$ Prior distribution

$KL(p\|q) \triangleq \mathbb{E}_p[\log(p/q)]$

Regressional-MedLDA

Side information

Numerical rate of a document in the corpus

Regression case

Y is the response response

$$\hat{y} \triangleq \mathbb{E}[Y|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2] = \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2]$$

Take expectation with respect to $q(\boldsymbol{\eta}, \mathbf{z})$

let

$$q(\boldsymbol{\eta}, \mathbf{z}) = \int_{\boldsymbol{\theta}} q(\boldsymbol{\eta}) q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\eta})$$

and

$q(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ be the approximation to $p(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2, \mathbf{y}, \mathbf{W})$

Upper bound of negative log-likelihood

$$\begin{aligned}\mathcal{L}^{bs}(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) &\triangleq -\mathbb{E}_q[\log p(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)] - \mathcal{H}(q(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\})) \\ &= KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + \mathbb{E}_{q(\boldsymbol{\eta})}[\mathcal{L}^s].\end{aligned}$$

Regressional-MedLDA

Objective function

$$\text{P2}(\text{MedLDA}^r): \min_{q,\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2,\boldsymbol{\xi},\boldsymbol{\xi}^\star} \mathbb{E}_{q(\boldsymbol{\eta})}[\mathcal{L}^s(q;\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2)] + KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\forall d,\ \text{s.t.}: \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \le \epsilon + \xi_d, & (\mu_d) \\ -y_d + \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \le \epsilon + \xi_d^\star, & (\mu_d^\star) \\ \xi_d \ge 0, & (v_d) \\ \xi_d^\star \ge 0, & (v_d^\star) \end{cases}$$

$$\begin{aligned} \mathcal{L}^s(q;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\delta^2) &\triangleq -\mathbb{E}_q[\log p(\{\boldsymbol{\theta}_d,\mathbf{z}_d\},\mathbf{y},\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\delta^2)] - \mathcal{H}(q(\{\theta_d,\mathbf{z}_d\})) \\ &\ge -\log p(\mathbf{y},\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\delta^2). \end{aligned}$$

Classificational-MedLDA

Side information

Discrete label of a document in the corpus

Multi-class case

Y is discrete response and take value from $\mathcal{C} \triangleq \{1, 2, \cdots, J\}$

Linear discriminant function

$$F(y, \mathbf{z}, \boldsymbol{\eta}; \mathbf{w}) = \boldsymbol{\eta}_y^\top \bar{\mathbf{z}}$$

Effective discriminant function

$$F(y; \mathbf{w}) = \mathbb{E}[F(y, \mathbf{Z}, \boldsymbol{\eta}; \mathbf{w})] = \mathbb{E}[\boldsymbol{\eta}^\top \mathbf{f}(y, \bar{Z}) | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}]$$

Take expectation with respect to $q(\boldsymbol{\eta}, \mathbf{z})$

Prediction rule

$$\hat{y} = \underset{y \in \mathcal{C}}{\operatorname{argmax}} F(y; \mathbf{w}) = \underset{y \in \mathcal{C}}{\operatorname{argmax}} \mathbb{E}[\boldsymbol{\eta}^\top \mathbf{f}(y, \bar{Z}) | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}]$$

Classificational-MedLDA

Objective function

$$\text{P3(MedLDA}^{\text{c}}\text{)}: \quad \min_{q,q(\boldsymbol{\eta}),\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}} \quad \mathcal{L}^u(q;\boldsymbol{\alpha},\boldsymbol{\beta}) + KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) + \frac{C}{D}\sum_{d=1}^{D}\xi_d$$

$$\forall d,\ y \neq y_d,\ \text{s.t.}: \quad \begin{cases} \mathbb{E}[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(y)] \geq \Delta\ell_d(y) - \xi_d \\ \xi_d \geq 0, \end{cases}$$

$\Delta\ell_d(y)$ Cost function measure the different between prediction and the true label eg.(0/1)

$$\mathbb{E}[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(y)] = F(y_d;\mathbf{w}_d) - F(y;\mathbf{w}_d)$$

$$\begin{aligned} \mathcal{L}^u(q;\boldsymbol{\alpha},\boldsymbol{\beta}) &\triangleq -\mathbb{E}_q[\log p(\{\boldsymbol{\theta}_d,\mathbf{z}_d\},\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\beta})] - \mathcal{H}(q(\{\boldsymbol{\theta}_d,\mathbf{z}_d\})) \\ &\geq -\log p(\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\beta}), \end{aligned}$$

A general framework

Objective function

$$\text{P5(MedTM)}: \min_{q(\Upsilon,H),\Psi,\boldsymbol{\xi}} \mathbb{E}_{q(\Upsilon)}\Big[\mathcal{L}^t(q(H|\Upsilon);\Psi,\Upsilon)\Big] + KL(q(\Upsilon)\|p_0(\Upsilon)) + U(\boldsymbol{\xi})$$
$$\text{s.t.}: \quad q(\Upsilon, H) \text{ satisfies the expected margin constraints.}$$

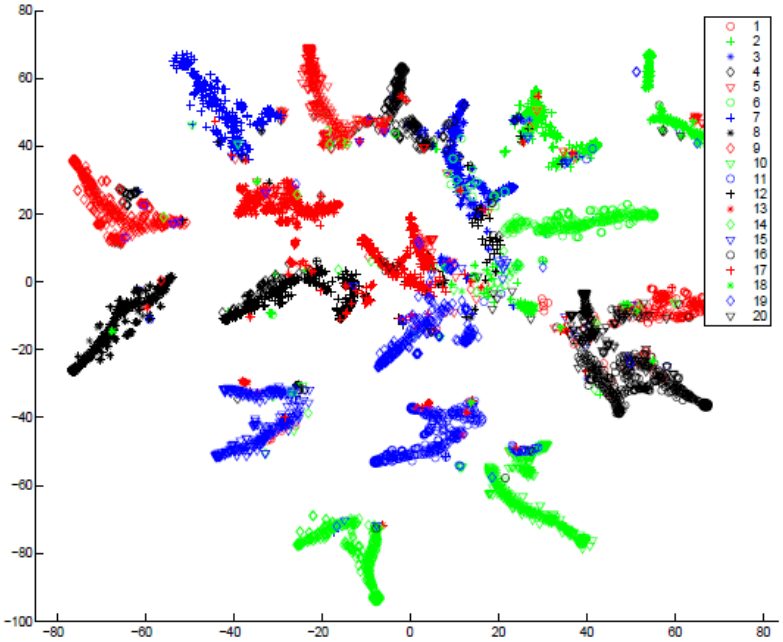
Υ	Denote the parameters of the model pertain to the prediction task	$\boldsymbol{\eta}$ in sLDA
H	Topic assignment and mixing variable	$\mathbf{z}$ and $\boldsymbol{\theta}$
Ψ	Parameter of the underlying topic model	$\boldsymbol{\alpha}$ and topics $\boldsymbol{\beta}$

Experiment

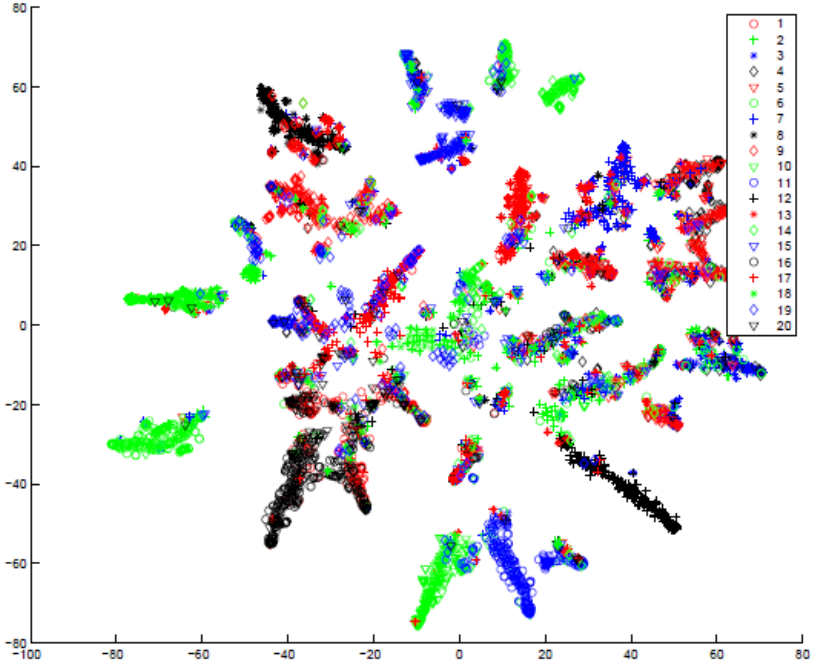
Dataset:

20 news group dataset, 20000 documents in 20 categories. Set the topic numbers to 110.

Embedding:



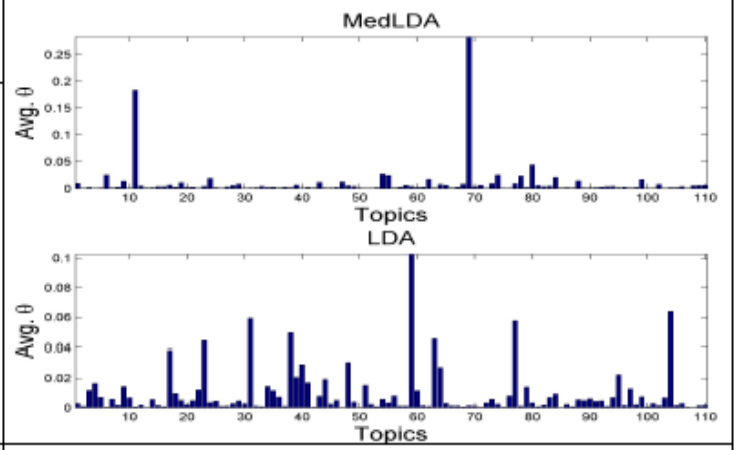
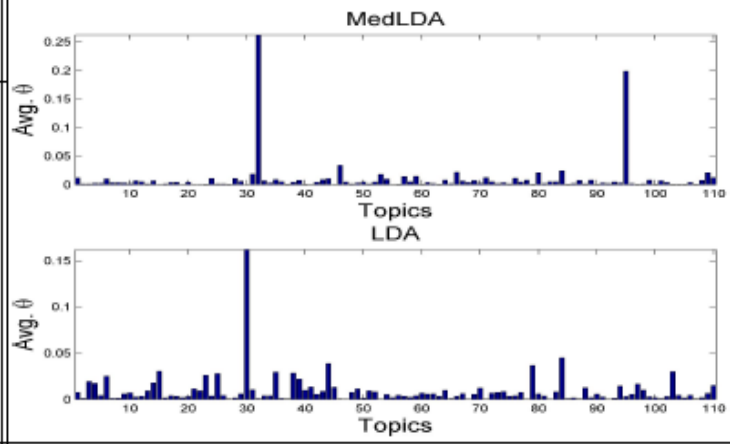
Med-LDA



LDA

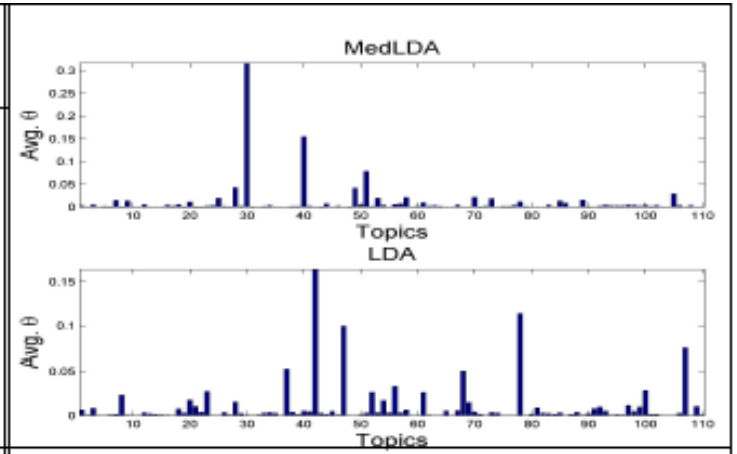
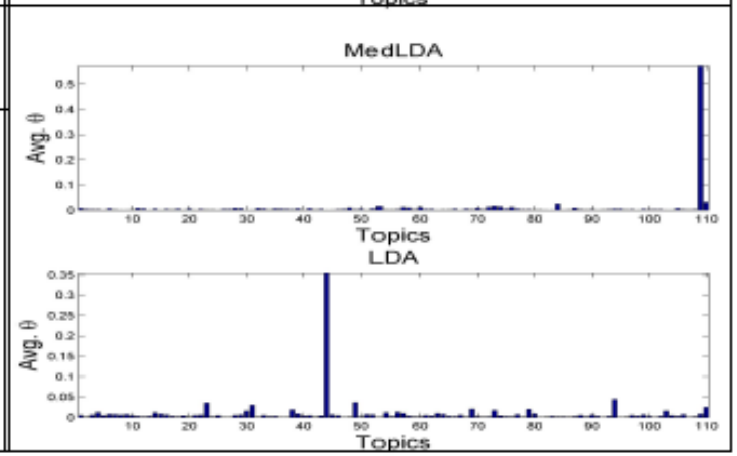
Experiment

Top topic

Class	MedLDA			LDA			Average θ per class
comp.graphics	T 69	T 11	T 80	T 59	T 104	T 31	
	image	graphics	db	image	ftp	card	
	jpeg	image	key	jpeg	pub	monitor	
	gif	data	chip	color	graphics	dos	
	file	ftp	encryption	file	mail	video	
	color	software	clipper	gif	version	apple	
	files	pub	system	images	tar	windows	
	bit	mail	government	format	file	drivers	
	images	package	keys	bit	information	vga	
	format	fax	law	files	send	cards	
	program	images	escrow	display	server	graphics	
sci.electronics	T 32	T 95	T 46	T 30	T 84	T 44	
	ground	audio	source	power	water	sale	
	wire	output	rs	ground	energy	price	
	power	input	time	wire	air	offer	
	wiring	signal	john	circuit	nuclear	shipping	
	don	chip	cycle	supply	loop	sell	
	current	high	low	voltage	hot	interested	
	circuit	data	dixie	current	cold	mail	
	neutral	mhz	dog	wiring	cooling	condition	
	writes	time	weeks	signal	heat	email	
	work	good	face	cable	temperature	cd	

Experiment

Top topic

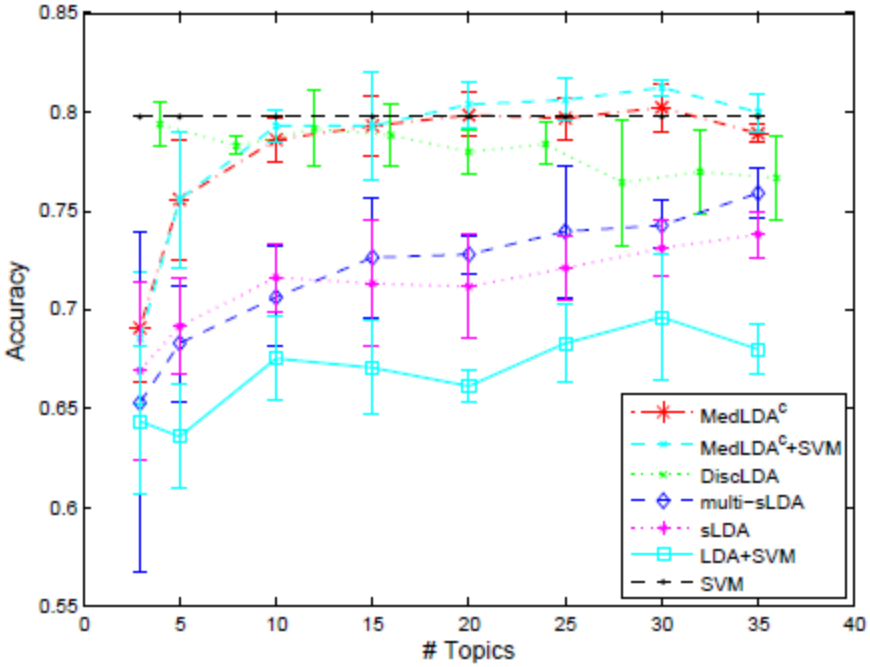
	T 30	T 40	T 51	T 42	T 78	T 47	
politics.mideast	israel israeli jews arab writes people article jewish state rights	turkish armenian armenians armenia people turks greek turkey government soviet	israel lebanese israeli lebanon people attacks soldiers villages peace writes	israel israeli peace writes article arab war lebanese lebanon people	jews jewish israel israeli arab people arabs center jew nazi	armenian turkish armenians armenia turks genocide russian soviet people muslim	
misc.forsale	T 109	T 110	T 84	T 44	T 94	T 49	
	sale price shipping offer mail condition interested sell email dos	drive scsi mb drives controller disk ide hard bus system	mac apple monitor bit mhz card video speed memory system	sale price offer shipping sell interested mail condition email cd	don mail call package writes send number ve hotel credit	drive scsi disk hard mb drives ide controller floppy system	

Experiment(classification)

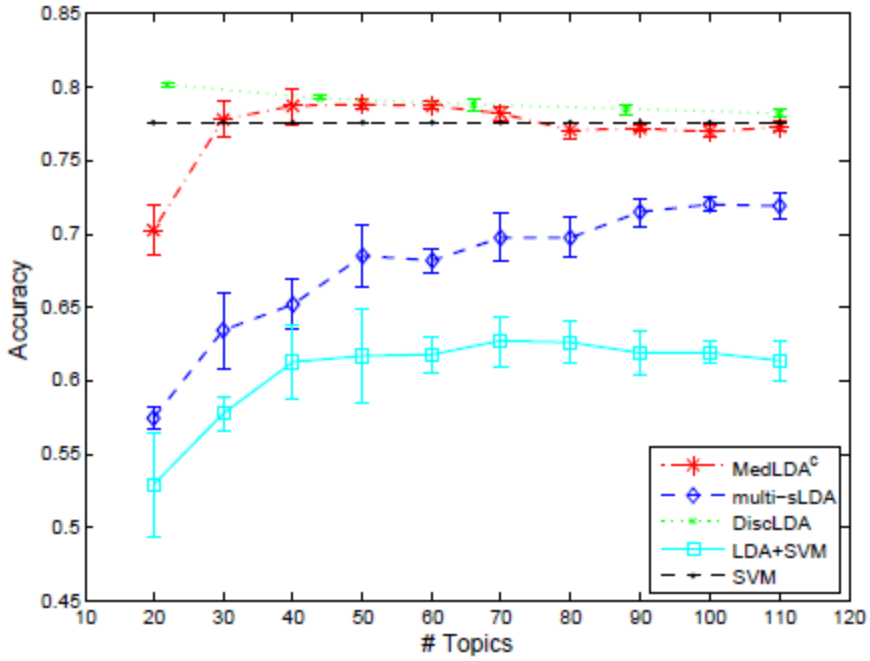
Dataset:

Binary classification:
Training: 856 documents with 480 and 376 over two categories

Multi-class classification (20 categories):
Training: 11269 documents
Testing:7505 documents



binary

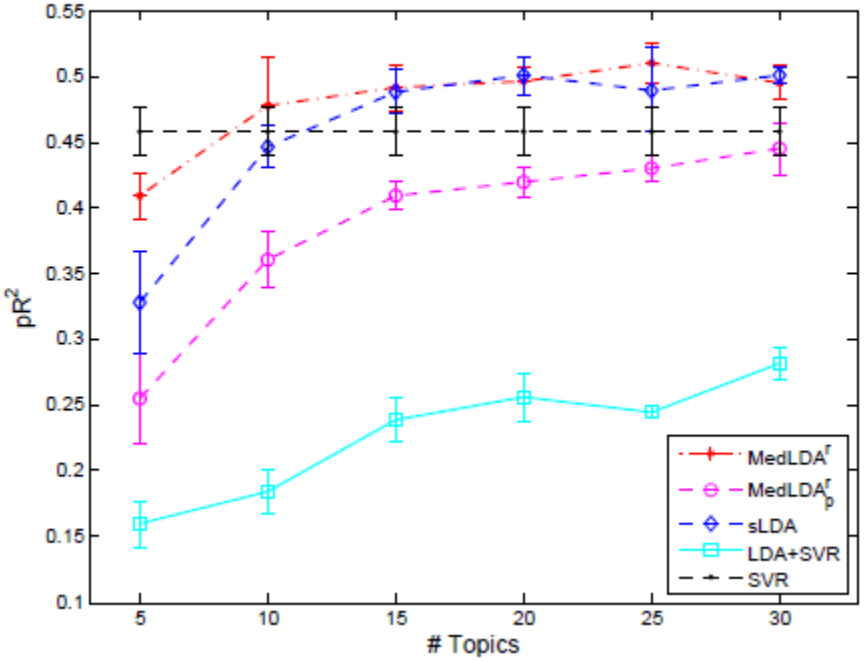


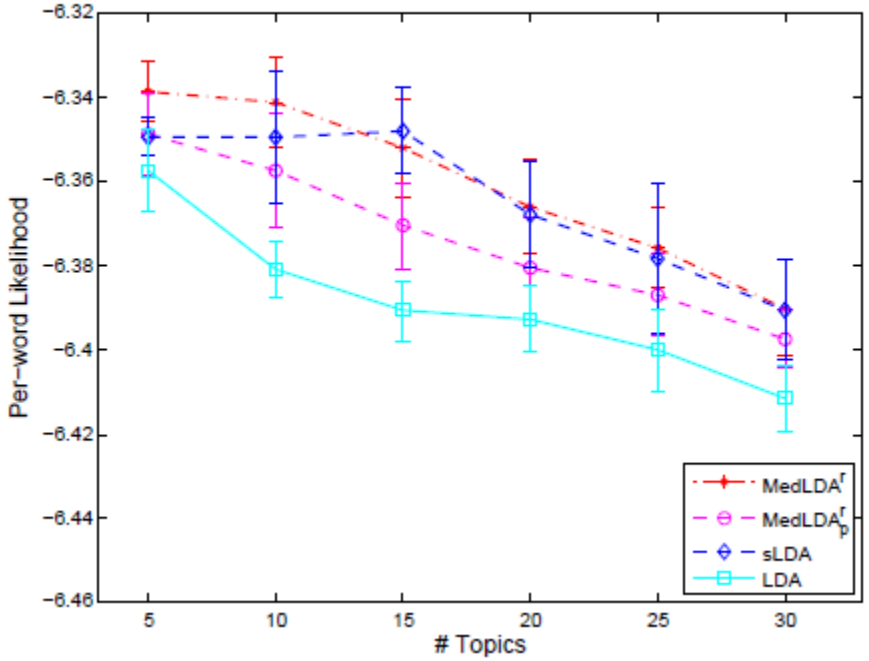
Multi-class

Experiment (regression)

Dataset:
Movie review data: 5006 documents

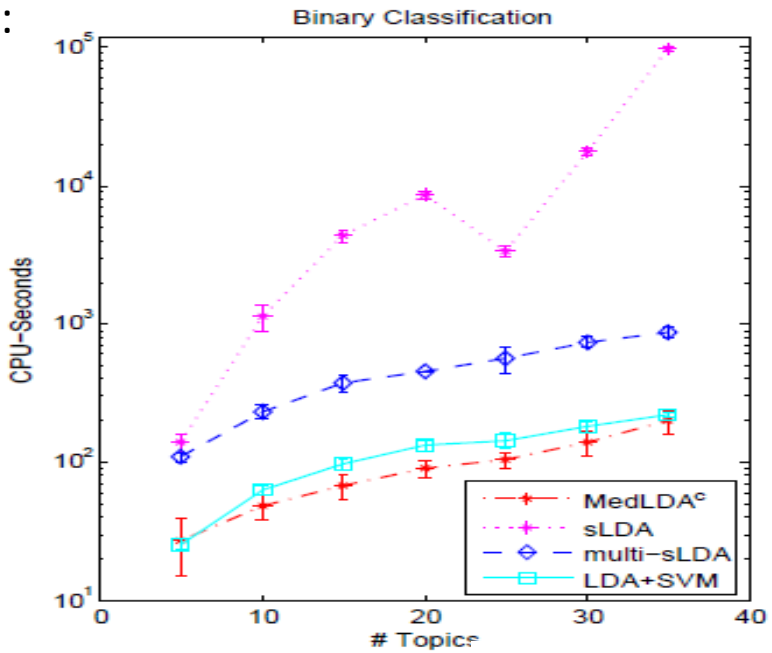
Measure criteria $$\mathrm{pR}^2 = 1 - \frac{\sum_{d=1}^{D}(y_d - \hat{y}_d)^2}{\sum_{d=1}^{D}(y_d - \bar{y})^2},$$

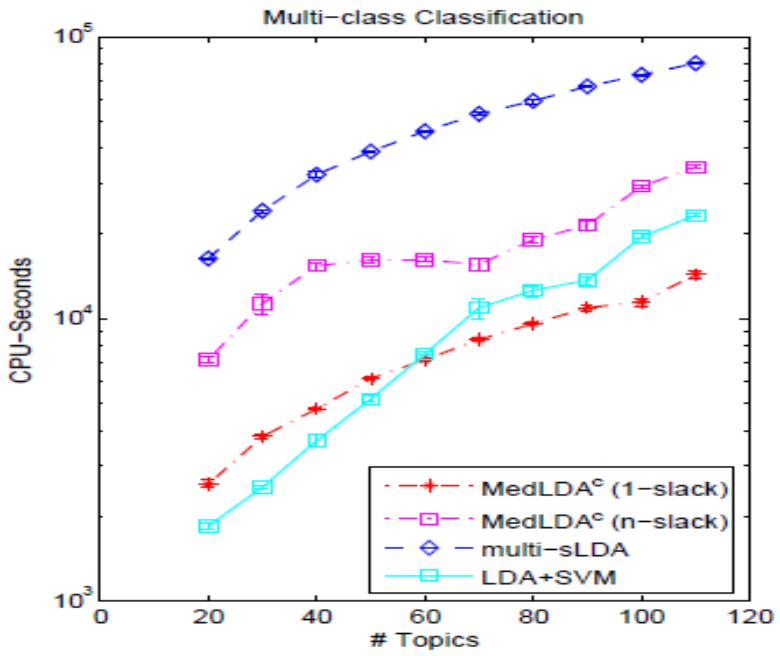




Experiment (time efficiency)

Training:





Testing: