Discriminative Topic Modeling with Logistic LDA

Iryna Korshunova[1*] Hanchen Xiong[2] Mateusz Fedoryszak[2] Lucas Theis[2]

[1]Ghent University [2]Twitter

[*]Work done at Twitter

Overview

Logistic LDA is a novel discriminative variant of latent Dirichlet allocation (LDA) which is easy to apply to arbitrary inputs, such as images or text embeddings.

Logistic LDA preserves LDA's extensibility and interpretability. In particular, it explicitly models item topics and group-level topic distributions, while integrating deep neural networks in a principled manner.

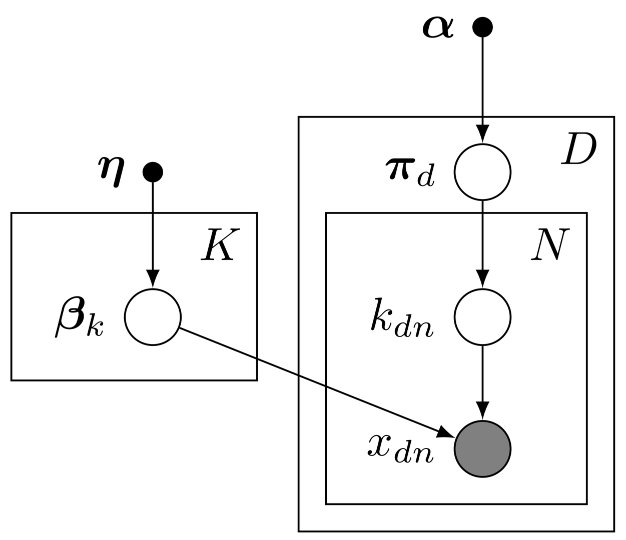
Among other desirable properties, **logistic LDA**:

- ✓ can be supervised, semi-supervised or unsupervised
- ✓ is scalable to large datasets
- ✓ can benefit from the vast literature on LDA
- ✓ applicable to a wide range of problems with group structure present in the data

Latent Dirichlet Allocation

- D - number of documents in a corpus
- N_d - number of words in a document d
- K - number of topics
- V - number of words in the vocabulary
- $\boldsymbol{x}_{dn}$ - n-th observed word in d-th document
- $\boldsymbol{k}_{dn}$ - latent topic of word $\boldsymbol{x}_{dn}$
- $\boldsymbol{\pi}_d$ - distribution over topics
- β - $K \times V$ matrix of topic-word distributions

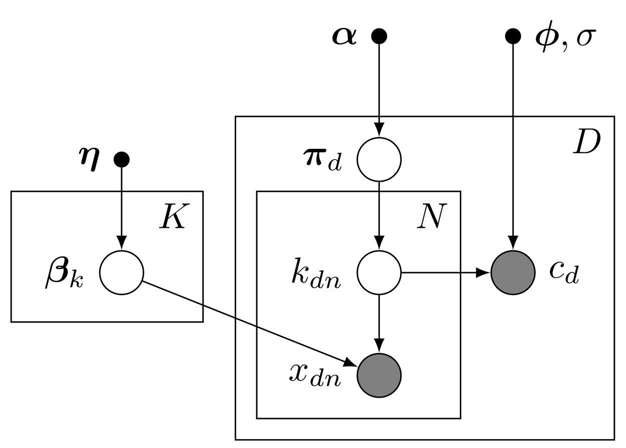
LDA:



Generative process:

1. Draw topic proportions $\boldsymbol{\pi}_d \sim \text{Dir}(\boldsymbol{\alpha})$
2. Draw topic-word distributions $\beta_k \sim \text{Dir}(\boldsymbol{\eta})$
3. For each word $\boldsymbol{x}_{dn}$:
   - 3.1 Draw a topic assignment $\boldsymbol{k}_{dn} \sim \text{Cat}(\boldsymbol{\pi}_d)$
   - 3.2 Draw a word $\boldsymbol{x}_{dn} \sim \text{Cat}(\beta^\top \boldsymbol{k}_{dn})$
4. In supervised LDA, draw a response variable:
   $c_d \sim \mathcal{N}(\boldsymbol{\phi}^\top(\frac{1}{N_d}\sum_n \boldsymbol{k}_{dn}), \sigma^2)$

sLDA [1]:'



Generative or Discriminative? [5]

Generative

$$p(\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{c} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$$
$$= p(\boldsymbol{c} \mid \boldsymbol{\pi})p(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\lambda})p(\boldsymbol{\theta}), \text{ with } \boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\lambda}\}$$

e.g. LDA, naive Bayes classifier, linear discriminant analysis, GMM

Discriminative

$$p(\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{c} \mid \boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x})$$

e.g. logistic regression, SVM, CRF

Logisitc LDA

$$p(\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{c}, \boldsymbol{\theta} \mid \boldsymbol{x})p(\boldsymbol{x})$$

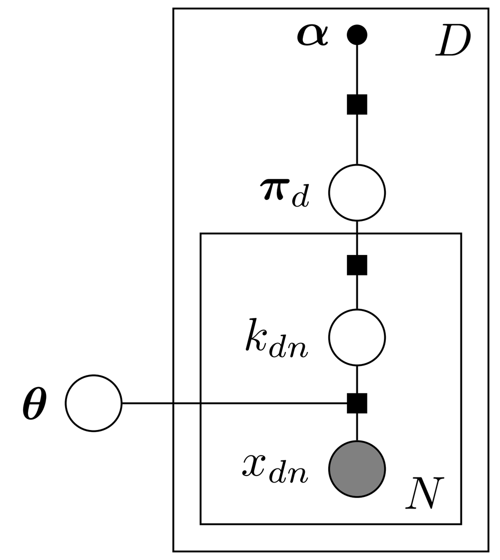
Alternative View of LDA

▶ We specify a set of full conditional probabilities:

$$p(\boldsymbol{\pi}_d \mid \boldsymbol{k}_d) = \text{Dir}\left(\boldsymbol{\pi}_d; \boldsymbol{\alpha} + \sum_n \boldsymbol{k}_{dn}\right)$$

$$p(\boldsymbol{k}_{dn} \mid \boldsymbol{x}_{dn}, \boldsymbol{\pi}_d, \boldsymbol{\theta}) = \boldsymbol{k}_{dn}^\top \text{softmax}(g(\boldsymbol{x}_{dn}, \boldsymbol{\theta}) + \ln \boldsymbol{\pi}_d)$$

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{k}) \propto \exp\left(r(\boldsymbol{\theta}) + \sum_{dn} \boldsymbol{k}_{dn}^\top g(\boldsymbol{x}_{dn}, \boldsymbol{\theta})\right)$$



▶ That result in a valid joint distribution:

$$p(\boldsymbol{\pi}, \boldsymbol{k}, \boldsymbol{\theta} \mid \boldsymbol{x}) \propto \exp\left((\boldsymbol{\alpha} - 1)^\top \sum_d \ln \boldsymbol{\pi}_d + \sum_{dn} \boldsymbol{k}_{dn}^\top \ln \boldsymbol{\pi}_d + \sum_{dn} \boldsymbol{k}_{dn}^\top g(\boldsymbol{x}_{dn}, \boldsymbol{\theta}) + r(\boldsymbol{\theta})\right)$$

▶ Where LDA is a special case:

$$g(\boldsymbol{x}_{dn}, \boldsymbol{\beta}) = \ln \boldsymbol{\beta}\ \boldsymbol{x}_{dn} \qquad r(\boldsymbol{\beta}) = (\boldsymbol{\eta} - 1)^\top \sum_k \ln \boldsymbol{\beta}_k \qquad \sum_j \boldsymbol{\beta}_{kj} = 1$$

Logistic LDA

$$g(\boldsymbol{x}_{dn}, \boldsymbol{\theta}) = \ln \text{softmax}\ f(\boldsymbol{x}_{dn}, \boldsymbol{\theta}) \qquad r(\boldsymbol{\theta}, \boldsymbol{x}) = \gamma \cdot \boldsymbol{1}^\top \ln \sum_{dn} \exp g(\boldsymbol{x}_{dn}, \boldsymbol{\theta})$$

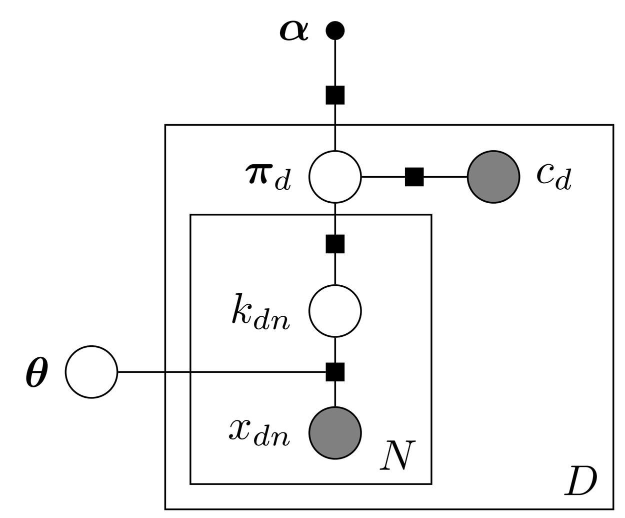
Supervised Logistic LDA

$$p(\boldsymbol{\pi}_d \mid \boldsymbol{k}_d, \boldsymbol{c}_d) = \text{Dir}\left(\boldsymbol{\pi}_d; \boldsymbol{\alpha} + \sum_n \boldsymbol{k}_{dn} + \lambda \boldsymbol{c}_d\right)$$

$$p(\boldsymbol{k}_{dn} \mid \boldsymbol{x}_{dn}, \boldsymbol{\pi}_d, \boldsymbol{\theta}) = \boldsymbol{k}_{dn}^\top \text{softmax}(g(\boldsymbol{x}_{dn}, \boldsymbol{\theta}) + \ln \boldsymbol{\pi}_d)$$

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{k}) \propto \exp\left(r(\boldsymbol{\theta}) + \sum_{dn} \boldsymbol{k}_{dn}^\top g(\boldsymbol{x}_{dn}, \boldsymbol{\theta})\right)$$

$$p(\boldsymbol{c}_d \mid \boldsymbol{\pi}_d) = \text{softmax}(\lambda \boldsymbol{c}_d^\top \ln \boldsymbol{\pi}_d)$$



Training and Inference

$$\min D_{\text{KL}}\left[q(\boldsymbol{\theta})\left(\prod_d q(\boldsymbol{c}_d)\right)\left(\prod_d q(\boldsymbol{\pi}_d)\right)\left(\prod_{dn} q(\boldsymbol{k}_{dn})\right) \mid\mid p(\boldsymbol{\pi}, \boldsymbol{k}, \boldsymbol{c}, \boldsymbol{\theta} \mid \boldsymbol{x})\right]$$

▶ Coordinate descent updates for variational parameters when $\boldsymbol{\theta}$ is fixed:

$$q(\boldsymbol{c}_d) = \boldsymbol{c}_d^\top \hat{\boldsymbol{p}}_d \qquad \hat{\boldsymbol{p}}_d = \text{softmax}(\lambda\psi(\hat{\boldsymbol{\alpha}}_d))$$
$$q(\boldsymbol{\pi}_d) = \text{Dir}(\boldsymbol{\pi}_d; \hat{\boldsymbol{\alpha}}_d) \qquad \hat{\boldsymbol{\alpha}}_d = \boldsymbol{\alpha} + \sum_n \hat{\boldsymbol{p}}_{dn} + \lambda \hat{\boldsymbol{p}}_d$$
$$q(\boldsymbol{k}_{dn}) = \boldsymbol{k}_{dn}^\top \hat{\boldsymbol{p}}_{dn} \qquad \hat{\boldsymbol{p}}_{dn} = \text{softmax}\left(f(\boldsymbol{x}_{dn}, \hat{\boldsymbol{\theta}}) + \psi(\hat{\boldsymbol{\alpha}}_d)\right)$$

▶ VI loss wrt. $\boldsymbol{\theta}$:

$$\ell(\hat{\boldsymbol{\theta}}) \approx -\sum_{dn} (\hat{\boldsymbol{p}}_{dn} + \gamma \cdot \hat{\boldsymbol{r}}_{dn})^\top g(\boldsymbol{x}_{dn}, \hat{\boldsymbol{\theta}})$$

▶ Alternative empirical loss when $\boldsymbol{c}_d$ is observed:

$$\ell(\hat{\boldsymbol{\theta}}) = -\sum_d \boldsymbol{c}_d^\top \ln \hat{\boldsymbol{p}}_d$$

Experiments: Twitter

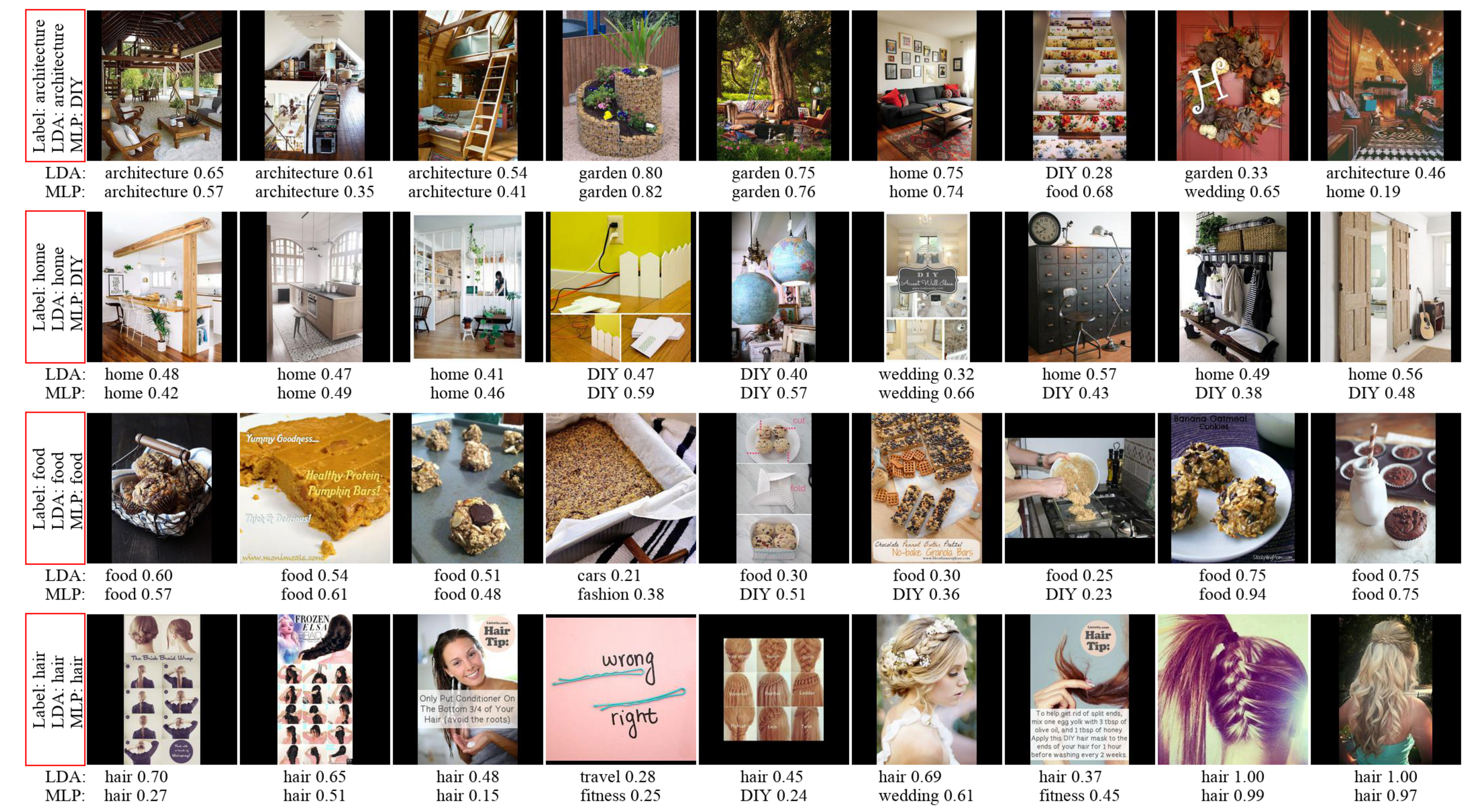
▶ Dataset of ∼4M tweets from ∼100K authors where some tweets and authors were annotated with one of 300 topics.

▶ In production: timeline filtering according to topics.

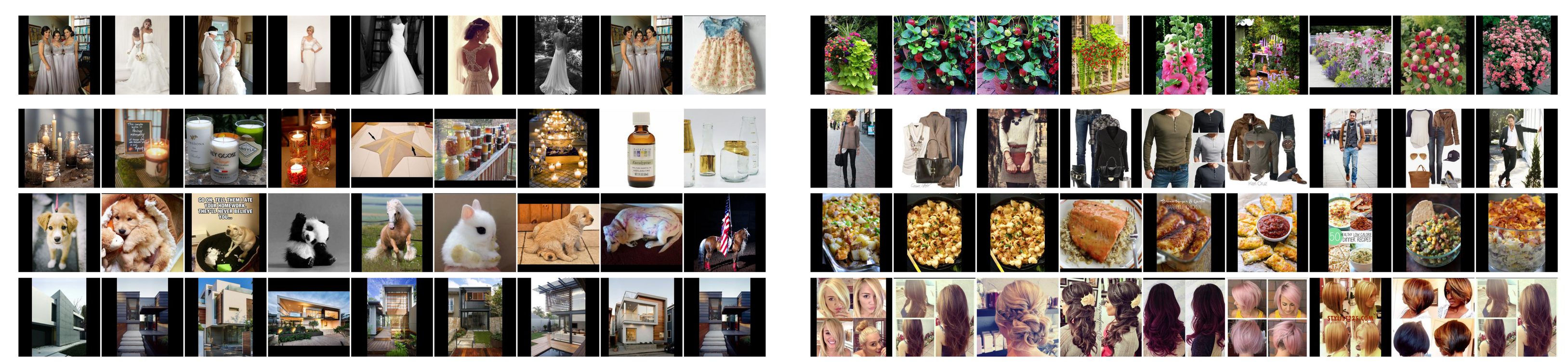
Model	Author	Tweet
MLP (individual)	26.6%	32.4%
MLP (majority)	35.0%	n/a
LDA	33.1%	25.4%
Logistic LDA	**38.7%**	**35.6%**

Experiments: Pinterest

Predictions for Pinterest boards and pins



Unsupervised topics



Experiments: 20-Newsgroups

Document classification accuracy

Logistic LDA	SVM [4]	LSTM [2]	SA-LSTM [2]	oh-2LSTMp [3]
84.4%	82.9%	82.0%	84.4%	86.5%

Unsupervised topics

1	bmw, motor, car, honda, motorcycle, auto, mg, engine, ford, bike
2	christianity, prophet, atheist, religion, holy, scripture, biblical, catholic, religious
3	spacecraft, orbit, probe, ship, satellite, rocket, surface, shipping, moon, launch
4	user, computer, microsoft, monitor,programmer, electronic, processing, data, app, systems
5	congress, administration, economic, accord, trade, criminal, seriously, fight, responsible, future

Bibliography

[1] D. Blei and J. McAuliffe. *Supervised Topic Models.* NIPS, 2008.
[2] A. Dai, Q. Le. *Semi-supervised Sequence Learning.* NIPS, 2015.
[3] R. Johnson, T.Zhang. *Supervised and Semi-supervised Text Categorization Using LSTM for Region Embeddings.* ICML, 2016.
[4] A. Cardoso-Cachopo. *Improving Methods for Single-label Text Categorization.* PhD thesis, Universidade Tecnica de Lisboa, 2007.
[5] J.Lasserre, C. Bishop *Generative or Discriminative? Getting the Best of Both Worlds.* Bayesian Statistics 8, 2007

Contact: lucas@theis.io, irene.korshunova@gmail.com Code: github.com/lucastheis/logistic_lda