

# Using Fisher Kernels from Topic Models for Dimensionality Reduction

Gaurav Chandalia and Matthew J. Beal

Department of Computer Science  
SUNY - University at Buffalo

December 14, 2006

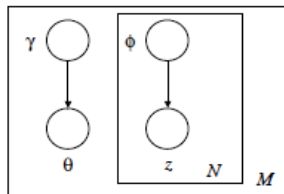
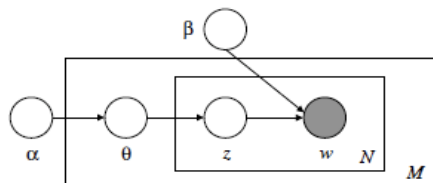
*cse@buffalo*

# Topic Models and Classification

- Objective
  - Learn a similarity metric from the data
  - Embed data points in a lower dimensional space
  - Combine kernels for classification
  
- Motivation
  - Technique to compare generative models for classification
  - Use topic models as feature extractors

# Latent Dirichlet Allocation (Blei et al., JMLR 2003)

- Represent documents as a set of topics



$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left( \prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left( \prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \quad (1)$$

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^N q(z_n|\phi_n) \quad (2)$$

# Fisher Kernel

- Take an *information geometric* perspective
- Effect of each data point on the model's parameters
- Invariant to reparameterization

$$K(\mathbf{w}_a, \mathbf{w}_b) \propto U_{\mathbf{w}_a}^T U_{\mathbf{w}_b} \quad (3)$$

$$U_{\mathbf{w}_a} = \nabla_{\theta} \log P(\mathbf{w}_a | \theta) \quad (4)$$

where  $w_a, w_b$  are data points

# Fisher Kernel for LDA

- $\mathcal{L}$  is the lower bound on the marginal likelihood

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) &= E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{z}|\theta)] + E_q[\log p(\mathbf{w}|\mathbf{z}, \beta)] \\ &\quad - E_q[\log q(\theta)] - E_q[\log q(\mathbf{z})]\end{aligned}\tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \Psi(\sum_{j=1}^k \alpha_j) - \Psi(\alpha_i) + \Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)\tag{6}$$

where  $\alpha_i$  and  $\gamma_i$  are the  $i^{\text{th}}$  component of the topic mixing proportion

- Overlap between the topics of the documents

$$\frac{\partial \mathcal{L}}{\partial \beta_{ij}} = \sum_{n=1}^N \frac{\phi_{ni} w_n^j}{\beta_{ij}}\tag{7}$$

where  $\beta_{ij}$  is the prob. of  $j^{\text{th}}$  word under the  $i^{\text{th}}$  component,  $\phi_{ni}$  is the variational posterior prob. that the  $n^{\text{th}}$  word is generated by the  $i^{\text{th}}$  topic

- Effect of common words generated from the same topic
- LDA is equivalent to pLSI for classification

# Experiments

- Dataset: 8000 documents from Reuters-21578
- Task: Binary Classification
  - EARN vs. NOT EARN
  - GRAIN vs. NOT GRAIN
  - U.S.A. vs. NOT U.S.A.
- Comparison
  - posterior Dirichlet parameters (topics of a document)
  - Fisher scores of a document obtained from derivatives w.r.t. the variational parameters

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)(\alpha_i + \sum_{n=1}^N \phi_{ni} - \gamma_i) - \Psi'(\sum_{j=1}^k \gamma_j) \sum_{j=1}^k (\alpha_j + \sum_{n=1}^N \phi_{nj} - \gamma_j) \quad (8)$$

# Results

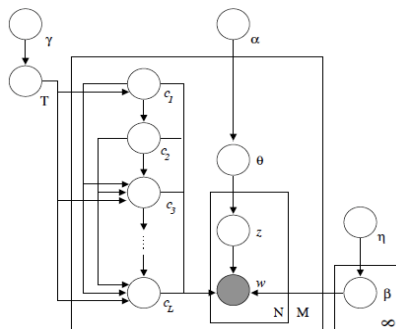
	Training Size	EARN vs. NOT EARN		GRAIN vs. NOT GRAIN		USA vs. NOT USA	
		LDA	Fisher LDA	LDA	Fisher LDA	LDA	Fisher LDA
Kernel evals/1000	400	168.2 ± 31.0	74.3 ± 14.8	12.0 ± 1.8	17.9 ± 3.8	69.5 ± 12.5	73.4 ± 24.8
	800	145.8 ± 21.6	70.0 ± 20.6	53.8 ± 5.3	18.3 ± 3.9	130.0 ± 26.5	198.7 ± 48.2
	1200	46.2 ± 5.74	42.5 ± 10.7	59.2 ± 7.6	17.9 ± 0.5	151.3 ± 22.2	306.0 ± 44.9
	1600	60.6 ± 11.5	117.5 ± 50.7	57.3 ± 7.6	22.5 ± 0.6	257.6 ± 40.6	384.4 ± 112.5
	2000	57.7 ± 1.42	50.6 ± 5.7	48.6 ± 3.1	26.8 ± 0.8	215.9 ± 46.2	265.4 ± 81.0
CPU runtime/s	400	233.5 ± 108.7	127.3 ± 98.0	0.7 ± 0.4	0.4 ± 0.1	42.0 ± 13.7	64.7 ± 53.2
	800	197.8 ± 130.2	26.2 ± 12.5	9.5 ± 3.5	0.3 ± 0.1	479.6 ± 269.6	29.1 ± 7.3
	1200	17.2 ± 2.5	28.9 ± 17.1	3.8 ± 0.9	0.3 ± 0.06	135.3 ± 53.7	41.6 ± 10.2
	1600	13.0 ± 1.3	103.4 ± 45.1	9.9 ± 2.5	1.5 ± 1.1	388.0 ± 310.2	135.9 ± 37.7
	2000	25.8 ± 2.9	10.1 ± 5.7	16.7 ± 5.8	2.4 ± 1.6	467.7 ± 252.7	407.9 ± 261.7
VC-dim	400	219.5 ± 69.3	37.0 ± 8.9	45.6 ± 18.5	2.8 ± 1.0	555.4 ± 94.4	232.1 ± 39.1
	800	64.5 ± 43.9	15.2 ± 9.1	5.7 ± 2.8	1.0 ± 0.002	526.0 ± 68.7	186.5 ± 32.5
	1200	2.1 ± 1.1	6.5 ± 5.5	1.2 ± 0.1	1.0 ± 0.009	395.2 ± 80.1	95.1 ± 28.2
	1600	1.9 ± 0.9	5.4 ± 4.4	1.0 ± 0.01	1.0 ± 0.009	190.6 ± 37.7	36.1 ± 12.7
	2000	1.0 ± 0.01	1.0 ± 0.007	1.1 ± 0.02	1.0 ± 0.001	93.5 ± 19.4	31.0 ± 19.0
Test accuracy	400	76.5 ± 0.2	77.5 ± 0.06	96.6 ± 0.1	96.9 ± 0.01	60.3 ± 0.2	60.6 ± 0.3
	800	77.4 ± 0.07	77.5 ± 0.04	96.8 ± 0.02	96.8 ± 0.02	60.6 ± 0.2	61.0 ± 0.2
	1200	77.5 ± 0.08	77.5 ± 0.09	96.8 ± 0.02	96.8 ± 0.02	61.1 ± 0.1	61.5 ± 0.1
	1600	77.5 ± 0.08	77.5 ± 0.09	96.8 ± 0.03	96.8 ± 0.03	61.4 ± 0.1	61.8 ± 0.08
	2000	77.4 ± 0.1	77.4 ± 0.1	96.8 ± 0.01	96.8 ± 0.01	61.7 ± 0.09	61.9 ± 0.06

## Results ...

- *Fisher LDA* is never less accurate than LDA
- Fisher LDA results in
  - Lesser kernel evaluations than LDA
  - Lesser training time than LDA
- Fisher LDA results in classifiers that have lower VC dimensions (upper bound) than those built from LDA
  - Large margin hyperplanes
  - Tighter bounds of generalization error

# hLDA (Blei et al., NIPS 2003)

- Uses a nested CRP prior to learn topic hierarchies
- Fisher kernel for hLDA
  - Compute the topic overlap at each level of the hierarchy



# Conclusions and Ongoing Work

- Fisher kernel results in classifiers with large margins thus ensuring good generalization
- Embeds features in a low dimensional space
- Preserves accuracy at a lower computational cost
  
- Obtain data dependent bounds
- Combine similarity functions of different models, for example, LDA and n-gram models

# References

- [BGJT03] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16*, 2003.
- [BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- [Elk05] Charles Elkan. Deriving tf-idf as a fisher kernel. In *12th International Conference on String Processing and Information Retrieval (SPIRE)*, 2005.
- [Hof00] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems 12*. The MIT Press, 2000.
- [JDH99] Tommi Jaakkola, Mark Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI, 1999.

# Lower Bound

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) &= E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{z}|\theta)] + E_q[\log p(\mathbf{w}|\mathbf{z}, \beta)] \\ &\quad - E_q[\log q(\theta)] - E_q[\log q(\mathbf{z})] \\ &= \log \Gamma(\sum_{j=1}^k \alpha_j) - \sum_{i=1}^k \log \Gamma(\alpha_i) + \sum_{i=1}^k (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\ &\quad + \sum_{n=1}^N \sum_{i=1}^k \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\ &\quad + \sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^V \phi_{ni} w_n^j \log \beta_{ij} \\ &\quad - \log \Gamma(\sum_{j=1}^k \gamma_j) + \sum_{i=1}^k \log \Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\ &\quad - \sum_{n=1}^N \sum_{i=1}^k \phi_{ni} \log \phi_{ni}\end{aligned}\tag{9}$$