

Using Fisher Kernels from Topic Models for Dimensionality Reduction

Gaurav Chandalia[†]

Matthew J. Beal[†]

gsc4@buffalo.edu

mbeal@cse.buffalo.edu

[†] Computer Science and Engineering, University at Buffalo, the State University of New York (SUNY), Buffalo, NY 14260 USA

Introduction: In this work we intend to address two issues. The first is to determine how topic models can be used as an effective tool for reducing the dimensionality of a feature set, and to analyze the performance when this reduced feature set is used for classification in a semi-supervised or unsupervised setting. Our primary motivation is to investigate whether such variations are an appropriate and concise representation for a document for classification purposes. We report in this abstract on our attempts to achieve this using an information geometric perspective, namely *Fisher kernels* [Jaakkola et al., 1999] of the topic models. The second issue that we intend to investigate in ongoing research is finding ways to combine the kernels obtained from different models. As an example in future work we intend to investigate the performance of models that incorporate both syntactic and semantic dependencies for classification. In particular we will analyze the Fisher kernel for latent Dirichlet Allocation [Blei et al., 2003b], LDA, which can capture the topical variations within a single document.

Fisher Kernel: The Fisher kernel makes use of the the first derivative of the log likelihood of a model with respect to its parameters θ , $U_{\mathbf{w}_a} = \nabla_{\theta} \log P(\mathbf{w}_a|\theta)$ where \mathbf{w}_a represents a document. This is denoted the *Fisher score*, which computes the effect of a new, or test, document on the model parameter θ . The Fisher Information matrix is defined by: $I = \mathcal{E}_{\mathbf{w}_a} [U_{\mathbf{w}_a} U_{\mathbf{w}_a}^{\top}]$, and the Fisher kernel defined by: $K(\mathbf{w}_a, \mathbf{w}_b) = U_{\mathbf{w}_a}^{\top} I^{-1} U_{\mathbf{w}_b}$. For computational reasons the information matrix is often approximated by the identity matrix, resulting in the kernel being the inner product $K(\mathbf{w}_a, \mathbf{w}_b) \propto U_{\mathbf{w}_a}^{\top} U_{\mathbf{w}_b}$. The Fisher kernel takes into account the posterior probability of the parameters and hence considers information obtained from the whole corpus.

Fisher Kernel for Latent Dirichlet Allocation: In this section we motivate the *Fisher kernel* for LDA. The intuition behind using a topic model for a document, or set of documents, is that the representation in terms of a reduced feature set can aide in a description of the document and therefore may improve the performance of discriminative classifiers; this is to be contrasted with the conventional practice of extracting as many features as possible from the document. Below we briefly show that we can interpret the features obtained from the Fisher kernel of LDA as being similar to those of probabilistic Latent Semantic Indexing [Hofmann, 2000]. We refer the reader to Blei et al. [2003b] for the description of the topic model LDA and the notation used. The model’s parameters and the variational parameters are denoted by $\{\alpha, \beta\}$ and $\{\phi, \gamma\}$, respectively, and we denote the *lower bound* on the marginal likelihood (the evidence) of a single document in terms of the variational parameters with \mathcal{L} . The derivative of \mathcal{L} with respect to the model’s parameters are functions of the variational parameters as follows:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \Psi(\sum_{j=1}^k \alpha_j) - \Psi(\alpha_i) + \Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j) \quad (1) \qquad \frac{\partial \mathcal{L}}{\partial \beta_{ij}} = \sum_{n=1}^N \frac{\phi_{ni} w_n^j}{\beta_{ij}} \quad (2)$$

where α_i and γ_i represent the i th component of the topic mixing proportion, out of k topics, β_{ij} the probability of j th word under the i th topic, and ϕ_{ni} is the variational posterior probability that the n th word was generated by the i th topic. The concatenation of these two derivatives comprises what we call the *variational Fisher score* for each document, from which can be computed the *variational Fisher kernel* between any two documents. We note that the exact Fisher kernel is intractable to compute for the very same reason the marginal likelihood in LDA is intractable.

Interpretation: The inner product of (1) evaluated for two documents computes the overlap between the topics responsible for generating the documents. Since we compute the similarity by considering the posterior probability of the topics, two documents will have high similarity value if their contexts are similar. Equation (2) computes the effect of the common words between the two documents on the same topic. However, since the Fisher score captures the effect of the word on the posterior probability of topics, if the context of the two documents is different then the common word will not increase the similarity between the documents. These results are similar to those obtained when computing the Fisher kernel for pLSI [Hofmann, 2000], showing that LDA is able to capture the features that are captured by pLSI. Moreover, we conjecture that the Fisher kernel for hierarchical LDA [Blei et al., 2003a] will be more informative than both LDA and pLSI as it considers a set of topics in a hierarchy before generating the word. Thus, intuitively we can compute the overlap between the topics at each level in the hierarchy of the two documents. This results in the documents being arranged in a hierarchical fashion based on the similarity between the features in the feature space. Also note that the parameter β is a topic specific multinomial distribution over words. We have not considered the case where we sample the multinomial vector from a Dirichlet distribution and then generate the words thus resulting in a topic specific Dirichlet Compound Multinomial (DCM) over words [Madsen et al., 2005]. Elkan [2005] have derived the Fisher kernel for DCM and showed that the empirically successful TF-IDF weighting scheme has its theoretical roots in the features derived from this kernel. Thus on one hand, LDA is able to capture *polysemy* and *synonymy* and on the other hand DCM captures the common notion of *word burstiness*. It would also be interesting to see how the kernels of these two models—which capture

different aspects of a document—can be combined to improve classification performance. The advantage here is that we are able to reduce the feature set through a principled approach without explicitly defining a similarity measure; instead the similarity function is induced by the generative model learnt for the data. We can also combine the feature sets obtained from different models by simply representing each feature vector as a list of Fisher scores from different models. Questions remain about how to weight the Fisher scores for this approach. As an example, we can model the semantic and syntactic dependencies of a document using LDA and the n-gram model and then combine their features using the corresponding Fisher scores.

Preliminary experiments: For this abstract we have performed experiments on document classification for the Reuters-21578 dataset in the same experimental setting outlined in Blei et al. [2003b]. The size of the dataset is 8000 documents. Blei et al. [2003b] used the posterior Dirichlet parameters γ as feature vectors to pass to an SVM, whereas we use as features the Fisher scores of the model with respect to these γ parameters, given in (3) below, to represent the reduced feature set for each document. We compared our results (which we will call *Fisher LDA*) to the results obtained from using only γ values as feature set for the following binary classification problems: EARN vs. NOT EARN, GRAIN vs. NOT GRAIN and U.S.A. vs. NOT U.S.A., given in Table 1 below.

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)(\alpha_i + \sum_{n=1}^N \phi_{ni} - \gamma_i) - \Psi'(\sum_{j=1}^k \gamma_j) \sum_{j=1}^k (\alpha_j + \sum_{n=1}^N \phi_{nj} - \gamma_j) \quad (3)$$

	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

Table 1: Comparison of the LDA and Fisher LDA in terms of number of kernel evaluations ($\times 1000$), training runtime in CPU seconds, estimated VC dimension of the classifier, and test accuracy.

Results & Conclusions: For each of the three data sets we ran 7-fold random replicates to obtain measures of performance, and report these as mean \pm standard error of the mean. We find that, in general, the number of kernel evaluations and CPU time for training the SVM is reduced for Fisher LDA compared to LDA (with the exception of some outliers in the USA data set). Furthermore, Fisher LDA is never less accurate than LDA, although one would need several more folds before we could be confident that Fisher LDA is statistically significantly better than LDA. What is most striking is that this equal or superior performance by Fisher LDA is achieved with classifiers that have substantially lower VC dimension than those built from LDA features, meaning the Fisher score results in a feature space having a much larger margin than that provided by simple posterior topic probabilities as done in Blei et al. [2003b]; having lower VC dimensions also gives smaller bounds on generalization error.

We have seen that the Fisher kernel obtained from a generative model results in a reduced set of dimensions and a small change in the feature values results in a large decrease in computational costs. Our next step is to determine how we can combine the Fisher scores obtained from models that characterize different features of a document at a lower computational cost and with little or no reduction in accuracy. Another direction that we want to pursue is to derive the Fisher kernel for hierarchical LDA [Blei et al., 2003a] to see how the similarity between the topics in the hierarchy affect the classification performance.

References

- 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*, 2003a.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003b. ISSN 1533-7928.
- Charles Elkan. Deriving TF-IDF as a fisher kernel. In *12th International Conference on String Processing and Information Retrieval (SPIRE)*, 2005.
- 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. ISBN 0-262-19450-3.
- 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. ISBN 1-57735-083-9.
- Rasmus Elsborg Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML)*. ACM, 2005. ISBN 1-59593-180-5.