Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Local Context Sparse Coding

Seungyeon Kim[*] and **Joonseok Lee**[*] and **Guy Lebanon**[†] and **Haesun Park**[*]
[*]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[†]Amazon, Seattle, WA, USA
{seungyeon.kim, jlee716}@gatech.edu, glebanon@gmail.com, hpark@cc.gatech.edu

Abstract

The n-gram model has been widely used to capture the local ordering of words, yet its exploding feature space often causes an estimation issue. This paper presents local context sparse coding (LCSC), a non-probabilistic topic model that effectively handles large feature spaces using sparse coding. In addition, it introduces a new concept of locality, local contexts, which provides a representation that can generate locally coherent topics and document representations. Our model efficiently finds topics and representations by applying greedy coordinate descent updates. The model is useful for discovering local topics and the semantic flow of a document, as well as constructing predictive models.

1 Introduction

Learning a representation that reflects word locality is important in a wide variety of text processing applications such as text categorization, information retrieval, or language model generation. The n-gram model, for example, is popular because of its simplicity and efficiency, which interprets a document as a collection of word sub-sequences. Specifically, it models a word given the previous $n - 1$ words: $p(w_i|w_{i-1}, \ldots, w_{i-n+1})$. The larger n is, the longer the contexts that the model can capture. A related approach is to model a symmetric window around a word $p(w_i|w_{i+1}, w_{i-1}, w_{i+2}, w_{i-2}, \ldots)$, as is done for example by Mikolov et al. (2013).

Lebanon, Mao, and Dillon (2007) extended local dependencies by applying different weights at each position of a document and summing up the word presence near a particular location. Specifically, that approach, named "locally weighted bag-of-words" (LOWBOW), uses a smoothing kernel to generate a smooth curve in the probability simplex that represents the temporal progression of the document. LOWBOW allows examining much longer-range dependencies than n-gram models, and it also allows tying word patterns to specific document locations. The bandwidth of the smoothing kernel captures the tradeoff between estimation bias and estimation variance. Our approach extends their work, but is different as it decouples local probabilities from their positions and it uses sparse coding to compress the parameter space.

Document models such as the n-gram and LOWBOW suffer from intrinsic sparsity, an inevitable consequence of capturing dependencies in sequences over a large vocabulary. The larger the dependency range, the harder it is to estimate the dependencies due to increased estimation variance. Specifically, the number of possible combinations of n consecutive words grows exponentially, making the number of observations for each combination extremely sparse, eventually causing not only computational difficulties but also a high estimation error. As a result, in many cases where data is limited, n-gram models with low n perform better than n-gram models with high values of n.

Neural probabilistic language models such as Bengio et al. (2006) are an attempt to handle this issue. They capture long term relations over a large vocabulary by using a parametric model that compresses the parameter space. Since the model estimates a compressed parameter vector rather than the exponentially growing n-gram counts, it is an effective way of capturing word dependencies that n-gram models cannot. On the other hand, probabilistic topic models such as Blei, Ng, and Jordan (2003) and matrix decomposition models (Deerwester et al. 1990; Lee and Seung 1999; Zhu and Xing 2011) estimate a compressed representation of the vocabulary, usually termed latent space or topics. Unlike the neural language model, these models are usually based on the bag-of-words representation or bigram features (Wallach 2006), limiting their potential to capture sequential word dependencies (though some recent extensions generalize topic models to sequential models - see Section 2).

By efficiently estimating sparse and compact representations of local dependencies, our model extends the work of Lebanon, Mao, and Dillon (2007) and Zhu and Xing (2011). We first define the notion of a *local context*, which is a conditional word probability given the word's location in the document. Similar to Lebanon, Mao, and Dillon (2007), we use a smoothing kernel to estimate the local context. Each kernel bandwidth examines a unique range of local resolutions. As noted earlier, because of the huge number of local contexts in our model, we apply a sparse-coding formulation to compress the space.

Our model has several benefits. First, by introducing rich local dependencies, it can generate highly discriminating

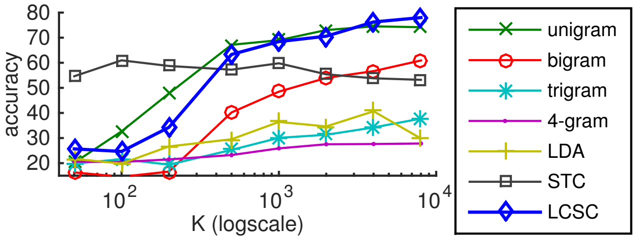

Figure 4: Test set classification accuracies with various sizes of dictionaries and methods

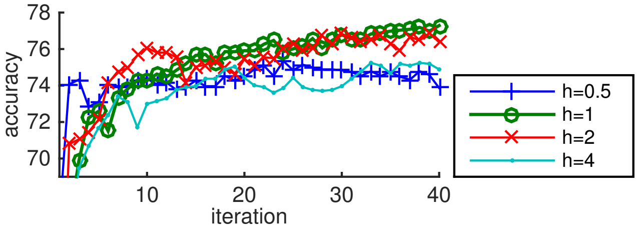

Figure 5: Test set classification accuracies of LCSC with various smoothing bandwidths

of size $|V| = 6328$. In the following two subsections, to examine the effect of parameters, we handle a subset of the dataset (5 classes, comp.*). In the last subsection, we evaluate overall performance on both the subset of the dataset and the whole dataset.

Effect of the Number of Topics (K) Figure 4 shows test set classification accuracies with various methods and sizes of dictionaries (from 50 to 8000). In the case of n-gram models, we selected the most frequent K features from the training set. For the other methods LDA[3], STC[4], and LCSC, we specify the size of a dictionary as a parameter. The bandwidth of LCSC was fixed to h=1, which covers about 7 words ($\pm 3h$). We tried a set of candidates for the remaining parameters and chose the best performing one (for example, $\lambda = \{10^{-4}, 10^{-2}, 10^{-1}, 0.5, 1\}$ for STC).

LCSC performs similar to unigram with small dictionaries, but it eventually achieves superior performance with a dictionary of sufficient size (from K=4000), that is, the performance of LCSC keeps improving even after $K{>}|V|$ (unigram model reaches maximum performance when $K{<}|V|$). STC performs well with relatively small dictionaries, but its maximum performance is not as good as other methods.

Figure 4 partially confirms Section 3.2. Bigram, trigram and 4-gram model do not perform well even with a large dictionary. It is because the number of features grows rapidly (bigram generates $23|V|$ features, trigram for $35|V|$, and 4-gram for $37|V|$) and thus will drastically lower the number of observations for each feature. On the contrary, even though LCSC covers approximately 7 neighboring words, it does not seem to suffer from sparsity and shows superior performance.

	n-gram	LDA	STC	LCSC	MedSTC
comp.*	74.53	40.67	60.97	**78.01**	77.70
*	74.10	34.43	61.14	**80.76**	79.81

Table 2: Comparision of test set classification accuracy for various methods on 5 classes (comp.*) and full 20 classes (*) of 20 newsgroup dataset

Effect of Bandwidth (h) Figure 5 shows test set classification accuracies of LCSC with various bandwidths while other parameters are fixed (K=4000, ρ=10^{-4}, λ=10^{-2}). The best performance was obtained at h=1. Using narrower bandwidth (h=0.5) led to faster convergence to poor performance, which is caused by lack of variability of local features. Using broader bandwidth (h=4) slowed down the convergence and ruined the performance, which is attributed to including unnecessary local dependencies for this task. The diverse results of various bandwidths confirms that locality features makes a notable difference in classification performance.

Comparision of Overall Performance We finally compare the overall performance of LCSC with other methods including a local-dependency model, n-gram, and unsupervised topic models: LDA and STC. We additionally included a top performing *supervised* topic model, MedSTC (Zhu and Xing 2011). Note, however, that MedSTC uses auxiliary supervised information (labeled data) during its topic learning, and cannot be directly compared to our method. We tried various sets of parameters and choose the best performing one (K: [50,...,8000], λ, ρ: [10^{-4},...,10^{-1}]). For n-gram models, we tried n: [1,...,4] and chose the best.

LCSC outperforms all other competitors on the subset as well as the full set (Table 2). The performance gain with respect to n-gram models shows that modeling long-range dependencies can be beneficial in classification. The better performance of LCSC compared to other methods including MedSTC (significant at p-value: 0.002) is notable since MedSTC directly optimizes for its discriminative performance whereas LCSC is a purely unsupervised coding method.

6 Summary

This paper presents a non-probabilistic topic model for local word distributions. Our model employed kernel smoothing to capture sequential information, which granted a flexible and efficient way to handle a wide range of local information. Our sparse-coding formulation leads to efficient training procedures, and a sparse representation that is locally coherent and has stronger discrimination capacity.

Acknowledgments

This work was supported in part by NSF Grant CCF-1348152 and DARPA XDATA Grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

[3] FastLDA: http://www-users.cs.umn.edu/~shan/mmnb_code.html

[4] http://www.ml-thu.net/~jun/stc.shtml

References

Bengio, Y.; Schwenk, H.; Senécal, J.; Morin, F.; and Gauvain, J. 2006. Neural probabilistic language models. In *Innovations in Machine Learning*. Springer. 137–186.

Blei, D., and Lafferty, J. 2006. Dynamic topic models. In *Proc. of the International Conference on Machine Learning*.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

Chen, H.; Branavan, S.; Barzilay, R.; and Karger, D. 2009. Content modeling using latent permutations. *Journal of Artificial Intelligence Research* 36(1):129–163.

Deerwester, S.; Dumais, S.; Landauer, T.; Furnas, G.; and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science* 41(6):391–407.

Dillon, J.; Mao, Y.; Lebanon, G.; and Zhang, J. 2007. Statistical translation, heat kernels, and expected distances. In *Uncertainty in Artificial Intelligence*, 93–100. AUAI Press.

Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; and Chandra, T. 2008. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proc of International Conference on Machine Learning (ICML)*.

Lebanon, G., and Zhao, Y. 2008. Local likelihood modeling of the concept drift phenomenon. In *Proc. of the 25th International Conference on Machine Learning*.

Lebanon, G.; Mao, Y.; and Dillon, J. 2007. The locally weighted bag of words framework for documents. *Journal of Machine Learning Research* 8:2405–2441.

Lebanon, G.; Zhao, Y.; and Zhao, Y. 2010. Modeling temporal text streams using the local multinomial model. *Electronic Journal of Statistics* 4.

Lebanon, G. 2005a. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory* 51(4):1283–1294.

Lebanon, G. 2005b. Information geometry, the embedding principle, and document classification. In *Proc. of the 2nd International Symposium on Information Geometry and its Applications*, 101–108.

Lee, D., and Seung, H. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791.

Lee, H.; Raina, R.; Teichman, A.; and Ng, A. 2009. Exponential family sparse coding with application to self-taught learning. In *International Joint Conferences on Artificial Intelligence*.

Li, Y., and Osher, S. 2009. Coordinate descent optimization for l1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Probl. Imaging* 3(3):487–503.

Mao, Y.; Dillon, J.; and Lebanon, G. 2007. Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics* 13(6):1208–1215.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *Workshop at International Conference on Learning Representations*.

Wallach, H. 2006. Topic modeling: beyond bag-of-words. In *Proc. of the International Conference on Machine Learning*.

Wang, C.; Blei, D.; and Heckerman, D. 2009. Continuous time dynamic topic models. In *Proc. of Uncertainty in Artificial Intelligence*.

Zhu, J., and Xing, E. 2011. Sparse topical coding. *In Proc. of Uncertainty in Artificial Intelligence (UAI)*.