E0:270 - Machine Learning - Adversarial Attacks on Latent Dirichlet Allocation

Abhishek Dugar [1] **Ishaan Sood** [2] **Vipul Rathore** [3]

Abstract

The authors intend to replicate the work by (Mei & Zhu, 2015a). In this paper, an adversarial attack on LDA is carried out by manipulating the corpus so as to make a particular word appear in a certain topic (at a higher rank, say top 10). This is done making minimal modifications in the data. This results in solving a bilevel optimization problem where attacker's risk function is minimized as high-level objective and learner's optimization for LDA as the lower-level objective. So far, the algorithm has been understood by the authors and the dataset to be used has been preprocessed. Authors intend to develop Python code (with or without a library) for the purpose.

1. Introduction

In the recent years text mining has proved to be a useful tool in a lot of applications. Topic modelling is a subset of text mining where one tries to extract the topics which forms the basis of the given corpus. Topic modelling can be used to form summaries, draw inferences, filter spam and the likes. Hence it is very likely that external forces will try to alter the model generated by the topic modelling tool especially in situations where the inferences from the document are used to drive decision or when topic modelling is used to filter spam.

LDA is a go to tool used in topic modelling. Given that the attacking party know the tools we use in topic modelling, we would like to discuss ways in which an external force might change the results of the model generated by minimally modifying the input to our corpus.

2. Motivation

Decisions are in general "data driven" nowadays. Data can be used to effect and affect changes. This is not just applicable to the corporate sector but to the politics as well. A recent example are the claims that user data from Facebook was used to sway the decisions of the voters in the US which might have contributed to the success of Trump's campaign. To be able to gain 'genuine' inferences is of importance now more than ever.

To be able to skew the inferences from data in ones own favor can be considered very powerful in the environment we live in. Small changes in the emails data can shift the particular email from spam to non-spam. By perturbing the online data one can shift something not so important into the political agenda. we need an in depth study on how much the data needs to be perturbed to make such changes in the output so that one can account for these and make adjustments accordingly.

3. Literature Review

LDA using variational inference techniques was invented by Blei et al (Blei et al., 2003). This paper describes the mathematical framework required for the model. Adversarial attacks on a variety of machine learning techniques have been studied extensively eg.(Nelson et al., 2008),(Mei & Zhu, 2015b) etc. However to the best of our knowledge, the work by (Mei & Zhu, 2015a), which we intend to reproduce, is the only existing work for attack on LDA. Further, to understand variational inference and KKT conditions, authors have referred to (Christopher, 2016) extensively

4. Model Description

4.1. LDA Variational inference

LDA is a generative model for K topics. k^{th} topic is mutinomial ψ_k with $\psi_k \sim dirichlet(\beta)$ or $Dir(\beta)$.

Each document d has a topic proportion multinomial θ_d with distribution $Dir(\alpha)$.The posterior of interest is $p(\psi\ ,\theta\ ,Z\ |W\ ,\alpha\ ,\beta)$.We estimate this using variational methods.

[1] [2] [3]Department of Computational and Data Sciences, Indian Insitute of Science, Bangalore. Correspondence to: Abhishek Dugar <dugarab@gmail.com>, Ishaan Sood <ishaansood@iisc.ac.in>, Vipul Rahore <vipulrathore@iisc.ac.in>.

Midterm Project Report, E0:270, Machine Learning, 2018.

Our objective function for the problem is a bilevel objective function given by - $min_{\mathbf{M}\ \epsilon\ M,\mu(M)}(R_A(\psi(M), \psi\ ^*))$,s.t. $\psi(M)$ are LDA topics learnt from **M**, where **M** is the modified training corpus and can belong to a specific set M of modified corpus.

4.2. The design choice of R_A functions and M sets

R_A is a loss function between the learned topic distribution $\psi(M)$ and the distribution that attacker wants the user to learn $\psi\ ^*$. We have 2 options for this loss function -

- l2 attacker risk function - Calculates the sum of the squares of the difference of the topics.
- ϵ -insensitive l2 attacker risk function defined as: $R_{A,\epsilon-l2}(\psi(\eta(M))\ \ ,\psi^*) = (1/2) * \Sigma_k\Sigma_v((\ |\psi(\eta(M))_{kv})) - \psi^*_{kv}\ | - \epsilon)\ _+)^2$

We define a M set in which the l_1 distance between the original matrix M_0 and the manipulated **M** is within a total change limit of L and the l_1 distance of each row is within a per-document change limit L_d.The small attacks are

$$\begin{aligned} M = \{\mathbf{M}\epsilon R^{D\times V} : ||\mathbf{M_0} - \mathbf{M}\ ||_1 <= L \\ \wedge\forall d : ||\mathbf{M_{0,d},_.} - \mathbf{M_{d},_.}||_1 <= L_d\} \end{aligned} \tag{1}$$

4.3. Descent Method

We use descent method to update **M** as

$$\mathbf{M}^{(\mathbf{t})} = Proj_M[\mathbf{M}^{(t-1)} - \lambda_t\nabla_M R_{A,\epsilon-l2}(\psi(\eta(M)), \psi^*)] \tag{2}$$

We can calculate $\nabla_M R_{A,\epsilon-l2}(\psi(\eta(M))\ , \psi^*)$ using chain rule.

5. Dataset description

As used in the paper (), we'll use the CONG dataset (Thomas et al., 2006). This contains floor-debate transcripts from the United States House of Representatives in 2005. We also intend to use the NIPS conference papers dataset consisting of abstracts, authors and texts of research papers presented in NIPS 2000 through 2012, downloaded from http://www.cs.toronto.edu/ roweis/.

Number of documents	6362
Average words per document	118
Size of the vocabulary	15547

Table 1. CONG dataset description

6. Preliminary Results

After some preprocessing, LDA was run on the CONG dataset and one of the topic contained top 10 words as displayed in fig.1. For experiment we intend to insert words like 'solar' in the top 20 words of this topic. The code can be found at https://bit.ly/2GuOo5C

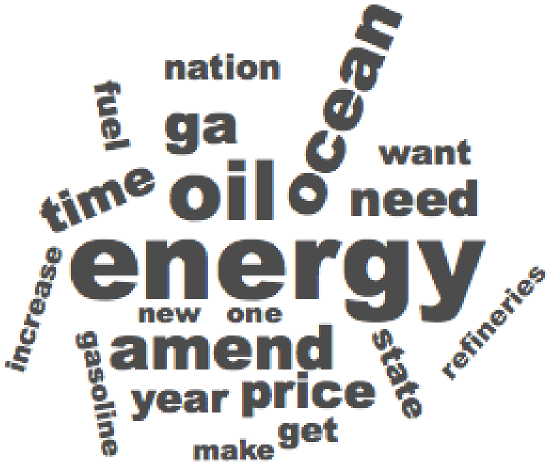

Figure 1. Top 20 words in the topic Energy

7. Future Work

To extend the current work, authors intend to explore deep-learning based topic modelling tools such as deep exponential families (Ranganath et al., 2014) and carry out a similar adversarial attack on them.

References

Blei, David M, Edu, Blei@cs Berkeley, Ng, Andrew Y, Edu, Ang@cs Stanford, Jordan, Michael I, and Edu, Jordan@cs Berkeley. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 15324435. doi: 10.1162/jmlr.2003.3.4-5.993.

Christopher, M Bishop. *PATTERN RECOGNITION AND MACHINE LEARNING.* Springer-Verlag New York, 2016.

Mei, Shike and Zhu, Xiaojin. The security of latent Dirichlet allocation. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 38:681–689, 2015a. ISSN 15337928.

Mei, Shike and Zhu, Xiaojin. Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2871–2877, 2015b.

Nelson, Blaine, Barreno, Marco, Chi, Fuching Jack, Joseph, Anthony D., Rubinstein, Benjamin I.P., Saini, Udam, Sutton, Charles, Tygar, J. D., and Xia, Kai. Exploiting machine learning to subvert your spam filter. *In Proceedings of the First Workshop on Large-scale Exploits and Emerging Threats (LEET)*, pp. Article 7, 2008.

Ranganath, Rajesh, Tang, Linpeng, Charlin, Laurent, and Blei, David M. Deep Exponential Families. *arXiv:1411.2581v1*, nov 2014. ISSN 15337928.

Thomas, Matt, Pang, Bo, and Lee, Lillian. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pp. 327–335, 2006.