Interpretable Machine Learning: Lessons from Topic Modeling

Michael J. Paul
University of Colorado
Boulder, CO 80309, USA
mpaul@colorado.edu

Abstract

This paper examines how the topic modeling community has characterized interpretability, and discusses how ideas used in topic modeling could be used to make other types of machine learning more interpretable. Interpretability is discussed both from the perspective of evaluation ("how interpretable is this model?") and training ("how can we make this model more interpretable?") in machine learning.

Author Keywords

interpretable machine learning; topic modeling; regression; coherence

ACM Classification Keywords

I.2.6 [Learning]: Concept learning, Parameter learning; H.1.2 [User/Machine Systems]

Introduction

Human interpretability is increasingly becoming recognized as an important property in machine learning models, but the machine learning community at large currently lacks standards for measuring interpretability, and there is not a clear path forward toward improvement.

One area of machine learning that has long focused on creating interpretable models is **topic modeling**. While topic models can be used for many purposes, they are often valued for their interpretability to humans, and as such, topic modeling researchers have proposed a number of methods for improving and evaluating the interpretability of topic models. This abstract considers how ideas from topic modeling could be applied to machine learning more generally. In particular, we will consider:

- Why should we care about the interpretability of machine learning models?
- How can interpretability be used as a criterion for **evaluation** as well as an objective for **training**?
- How should interpretability be defined, either as a function of **human feedback** or with **automated metrics**?

This abstract surveys how these issues have been addressed in topic modeling, and considers how these ideas can inform other areas of machine learning.

Why Does Interpretability Matter?

Usable Models: Let Machines Help Humans

Machine learning has the potential to aid advancements in a variety of domains such as medicine, finance, and the humanities. However, experts in these domains are unlikely to widely adopt machine learning tools if they do not understand or trust these tools [22, 5]. For example, topic modeling research has found that domain users are unlikely to trust topic models if some of the topics look incoherent or do not meet prior expectations [16, 23].

Moreover, as algorithms bleed into everyday life, it is crucial that people can understand what these algorithms are doing. For example, the European Union is considering legislation that would require algorithmic decision-making to be more transparent and explainable [7]. As pointed out by [10], this is a challenging task when using complex models such as deep neural networks.

Better Models: Let Humans Help Machines

Another motivating factor for interpretable models is that it is arguably easier to improve a model if a human can understand how it works. Traditional predictive metrics for evaluating machine learning models are not always sufficient for identifying if a model has been overfit or otherwise has “quirks” that will not generalize well, especially when only a limited amount of data is available. On the other hand, humans can often evaluate whether a model makes sense and can potentially identify bad models. This insight has led to an increased interest in interactive machine learning systems that use human feedback [1].

As an example of the need for human feedback, consider Google Flu Trends [9], a system that automatically estimates weekly influenza prevalence based on how many users are searching about the flu in a given week. During development of the system, seemingly irrelevant search queries were found to have very high correlation with influenza trends (including search terms related to winter holidays and high school basketball), simply because these events have similar seasonal patterns as the flu [8]. The result was a system that has been called “part flu detector, part winter detector” [14], rather than a system that truly modeled the concept of flu. Yet, identifying these spurious correlations is easy for a human—and indeed, Google’s final model used a smaller set of queries that were manually selected by people for relevance. But such errors can only be identified if a human can easily interpret how these features are being used by the model.

Interpretability in Topic Models

Topic models [3] are probabilistic models that represent text documents as mixtures of underlying “topics”. Each document is modeled as having a probability distribution over topics, while each topic is associated with a distribution over

Coherent	
space	health
earth	disease
moon	aids
science	virus
scientist	vaccine
light	infection
nasa	hiv
mission	cases
planet	infected
mars	asthma

Incoherent	
dog	king
moment	bond
hand	berry
face	bill
love	ray
self	rate
eye	james
turn	treas
young	byrd
character	key

Table 1: Examples (from [17]) of coherent and incoherent topics learned from a news corpus, as judged by humans.

words. Topic models are a form of unsupervised machine learning, in that the topics and mixture parameters are unknown and are inferred solely from the data. Even though the learning is unsupervised, the inferred model parameters often make sense to humans.

Topics are usually visually represented by the 10 or 20 most probable words in each topic, and so topics are typically interpreted as word clusters. Humans often judge topics based on whether the words in each topic cluster form interpretable concepts [6]. A number of approaches have been explored for evaluating how well topics are interpreted as coherent concepts, as well as approaches for training topic models in ways that are more satisfying to end users.

Evaluation

While the most common means of evaluating topic models involve measuring the performance at predictive tasks, such as the likelihood of held-out data [24], a number of methods have been proposed for evaluating how humans interpret topic models.

Human Feedback A number of topic modeling studies have measured topic quality by soliciting human feedback, sometimes by rating the quality of topics directly [15, 20], and other times by having humans complete tasks that indirectly measure interpretability [19]. In particular, **intrusion** tasks have been a popular way of measuring how well topics form coherent concepts [6]. For example, a person will be shown a number of words from a topic as well as one word from a different topic, and be asked to identify the out-of-place word: this task is easy with coherent topics and hard with incoherent topics, so performance at this task is indicative of topic quality.

Automated Metrics More recent research has investigated whether the quality of topics can be computed automatically. Metrics to measure the **coherence** of topics were introduced by [17] and [16], and later improved upon and generalized by others [13, 21]. While there is some variation among the various metrics, they are all based on word co-occurrence statistics. The idea is that all pairs of words within a topic should be related to each other, and co-occurrence metrics are a simple way of estimating word relatedness. For example, since words like *cough* and *fever* tend to occur together in the same documents, they would be considered coherent if they appeared together in a topic, whereas words like *cough* and *beehive* are unlikely to occur together, and are thus considered incoherent.

Training

While interpretability has primarily been used as a criterion for evaluation, there has also been some work on training and optimizing topic models in ways that will be more interpretable, both automatically and with human intervention.

Human Feedback Work has been done to design topic models that can incorporate human preferences, such as specifying which words should or should not appear in the same topic [2], or using seed words to guide topics [12, 18]. There has even been research on incorporating such preferences interactively, allowing users to change topics during the inference procedure [11].

Objective Functions Some research has modified the topic model likelihood objective in a way that encourages the formation of coherent topics. For example, after developing co-occurrence-based metrics for coherence, [16] then developed a modified topic model so that co-occurring words are likely to be in the same topic. Word co-occurrences have been used in other ways to improve interpretability, such as encouraging topics to be diverse [25].

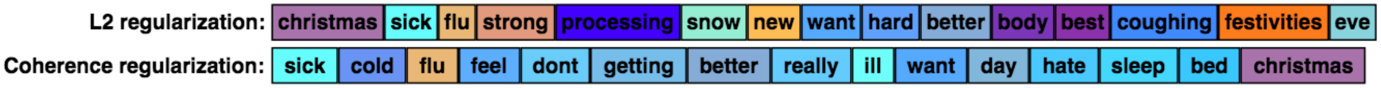

Figure 1: The 15 predictors with the highest coefficients (from left to right) in two bag-of-words linear regression models trained to estimate weekly flu prevalence based on word counts in health-related tweets (from [4]). The top row is a standard L2-regularized model, while the bottom row uses an experimental regularizer that encourages the regression coefficients to have similar covariance as the words in the data.

Lessons for Machine Learning

Many of the approaches to evaluating and improving interpretability in topic models could potentially apply to other areas of machine learning. In particular, the idea of **coherence** could be a desirable property in many classification and regression models, and it is worth considering this as a criterion in machine learning beyond topic modeling. While interpretability certainly depends on factors other than coherence, coherence plays an important role, and adopting a metric such as coherence would provide a common standard by which to compare different models for different tasks across the machine learning community.

As an example of how coherence can improve other machine learning models, consider a preliminary experiment on the task of estimating flu prevalence from user web activity, as described earlier. Figure 1 shows the most predictive features learned from training a bag-of-words linear regression model to estimate weekly flu counts from Twitter messages. The top row shows results when using a standard regression model with L2 regularization. Because the flu happened to spike during the week of Christmas during this particular season, the top predictor is not *sick* or *flu*, but *christmas*—a word that does not coherently fit into the concept of flu.

In contrast, the bottom row of the figure shows results when using a novel regularization approach that encourages words with similar co-occurrence patterns to have similar regression parameters. Specifically, the parameters have a multivariate normal distribution as the prior which encodes covariance between parameters:

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{1}$$

where $\boldsymbol{\eta}$ is the vector of regression coefficients, and $\boldsymbol{\Sigma}$ is the sample covariance of the words in the corpus. This prior encourages words that are positively (or negatively) correlated to have similar (or dissimilar) coefficients. As seen in Figure 1, when using this prior, *christmas* is no longer a strong predictor, and *sick* is the top feature.

While much more research is needed to understand whether and under what circumstances coherence may be useful for general machine learning tasks, we can begin by taking advantage of the large body of interpretability research that has already been done in the topic modeling community.

References

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* (2014).

[2] D. Andrzejewski, X. Zhu, and M. Craven. 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *ICML*.

[3] David M. Blei and John D. Lafferty. 2009. *Topic models*. Chapman & Hall/CRC.

[4] D.A. Broniatowski, M.J. Paul, and M. Dredze. 2013. National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE* 8, 12 (2013).

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *KDD*.

[6] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

[7] European Commission. 2012. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (2012). http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

[8] Samantha Cook, Corrie Conrad, Ashley L Fowlkes, and Matthew H Mohebbi. 2011. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE* 6, 8 (2011), e23610.

[9] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[10] E. Horvitz and D. Mulligan. 2015. Policy forum. Data, privacy, and the greater good. *Science* 349, 6245 (Jul 2015), 253–255.

[11] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith. 2013. Interactive Topic Modeling. *Machine Learning* 95 (2013), 423–469.

[12] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *EACL*.

[13] Han Jey Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *EACL*.

[14] D. Lazer, R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6167 (2014), 1203–1205.

[15] W. Li and A. McCallum. 2006. Pachinko Allocation: DAG-structured mixture models of topic correlations. In *International Conference on Machine Learning*.

[16] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. In *EMNLP*.

[17] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 100–108.

[18] M.J. Paul. 2015. *Topic Modeling with Structured Priors for Text-Driven Science*. Ph.D. Dissertation. Johns Hopkins University.

[19] Michael Paul and Mark Dredze. 2012. Factorial LDA: Sparse Multi-Dimensional Text Models. In *NIPS*.

[20] M. Paul and R. Girju. 2010. A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics. In *AAAI*.

[21] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *International Conference on Web Search and Data Mining (WSDM)*.

[22] Cynthia Rudin. 2014. Algorithms for Interpretable Machine Learning. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[23] E.M. Talley, D. Newman, D. Mimno, B.W. Herr II, H.M. Wallach, G.A.P.C. Burns, M. Leenders, and A. McCallum. 2011. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods* 8, 6 (2011), 443–444.

[24] H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. 2009. Evaluation methods for topic models. In *ICML*.

[25] James Y. Zou and Ryan P. Adams. 2012. Priors for Diversity in Generative Latent Variable Models. In *NIPS*.