**Supervised Topic Modeling using Hierarchical Dirichlet Process-based Inverse Regression: Experiments on E-Commerce Applications**

**Abstract:**

The proliferation of e-commerce calls for mining consumer preferences and opinions from user-generated text. To this end, topic models have been widely adopted to discover the underlying semantic themes (i.e., topics). Supervised topic models have emerged to leverage discovered topics for predicting the response of interest (e.g., product quality and sales). However, supervised topic modeling remains a challenging problem because of the need to prespecify the number of topics, the lack of predictive information in topics, and limited scalability. In this paper, we propose a novel supervised topic model, Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain as much predictive information as the original corpus. Moreover, we develop an efficient inference algorithm capable of examining large-scale corpora (millions of documents or more). Three experiments were conducted to evaluate the predictive performance over major e-commerce benchmark testbeds of online reviews. Overall, HDP-IR outperformed existing state-of-the-art supervised topic models. Particularly, retaining sufficient predictive information improved predictive R-squared by over 17.6 percent; having topic structure flexibility contributed to predictive R-squared by at least 4.1 percent. HDP-IR provides an important step for future study on user-generated texts from a topic perspective.

**Architecture:**

****

**Introduction:**

THE proliferation of e-commerce has given rise to a significant amount of user-generated text, which contains salient information about consumer preferences and opinion. Topic models are a major family of text analysis techniques for exploring the underlying semantic themes (i.e., topics) within textual data. However, prior research necessitates not only understanding the semantic themes but also integrating predictive analytics on variables of interest, such as customer sentiment, product quality, affect, and more. Standard topics models (e.g., LDA) are unsupervised and therefore incapable of making such predictions. To this end, the supervised topic modeling techniques have emerged, which can simultaneously discover the underlying semantic themes and leverage these themes for prediction. Both the discovered themes and the predicted response variables provide valuable insights about consumer preferences and opinions. Supervised topic models have a number of important e-commerce applications, including customer feedback assessment, online review evaluation, consumer sentiment analysis, product attributes mining, and customer preferences identification. However, supervised topic modeling remains a challenging problem. First, most supervised topic models require prespecifying the number of topics a priori. Such specification may result in model misspecification when the specified number of topics misrepresent the true underlying topic structure. For example, customer reviews for new products may contain unseen topics about new features. Prespecifying the number of topics inhibits the incorporation of such unseen topics, leading to unreliable topics and inaccurate predictions. Second, existing supervised topic models treat the proportion of topic mixtures as a reduced dimension representation of the original document and make predictions based on such representations. It is unclear whether these representations contain sufficient predictive information about the response. Statistically speaking, sufficiency entails that the reduced dimension representation preserves all the information from original documents for making predictions. The missing information in the supervised topic modeling process may diminish the prediction accuracy. Third, large text corpora often span several million documents, leaving many supervised topic models unscalable. Most supervised topic models adopt sampling-based inference algorithms, which require hundreds of iterations over each variable across all documents before convergence. Therefore, the scalability of these models is limited.

**Existing system:**

First, most supervised topic models require prespecifying the number of topics a priori. Such specification may result in model misspecification when the specified number of topics misrepresent the true underlying topic structure. For example, customer reviews for new products may contain unseen topics about new features. Prespecifying the number of topics inhibits the incorporation of such unseen topics, leading to unreliable topics and inaccurate predictions. Second, existing supervised topic models treat the proportion of topic mixtures as a reduced dimension representation of the original document and make predictions based on such representations. It is unclear whether these representations contain sufficient predictive information about the response. Statistically speaking, sufficiency entails that the reduced dimension representation preserves all the information from original documents for making predictions. The missing information in the supervised topic modeling process may diminish the prediction accuracy. Third, large text corpora often span several million documents, leaving many supervised topic models unscalable. Most supervised topic models adopt sampling-based inference algorithms, which require hundreds of iterations over each variable across all documents before convergenc. Therefore, the scalability of these models is limited.

**Proposed system:**

In this paper, we propose a novel supervised topic model called Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). Specifically, the Hierarchical Dirichlet Process (HDP) is a nonparametric topic modeling technique that allows for a flexible number of topics. Inverse Regression (IR) is a sufficient dimension reduction (SDR) technique that makes predictions with provably sufficient information. HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain statistically sufficient information for improved predictive performance. Moreover, we develop an efficient inference algorithm for model estimation that is capable of examining large-scale corpora with millions of documents. Evaluation of HDP-IR in comparison with the state-of-the-art baseline techniques reveals that both increasing the topic structure flexibility and using sufficient dimension reduction could improve the predictive performance on user-generated review text in e-commerce applications, and the proposed inference algorithm is highly effective in terms of its scalability. This paper is organized as follows. Provides a review of related work on major supervised topic models and identifies research gaps. Section 3 briefly introduces the background of HDP and IR. Details our proposed model and the algorithms for estimating the model and making predictions. It includes the experimental evaluation of the proposed model in comparison with the state-of-the-art baseline techniques. It Provides the conclusion and future directions.

**Modules:**

**Inference and Prediction:**

The traditional variational inference algorithm needs to perform coordinate ascent over both document-level variables for all documents (i.e., E-step) and corpus-level variables in each iteration (i.e., M-step). When the traditional variational algorithm examines large corpora containing hundreds of thousands of documents, the computational complexity associated with the E-step grows significantly. Based on stochastic optimization, SVI incorporates random subsampling into the E-step and then uses the resulting accumulated document-level sufficient statistics to optimize the corpus-level variables through natural gradient ascent.

**Inverse Regression:**

Inverse regression (IR) is a prominent SDR technique for textual data. Classical regression analysis focuses on estimating the conditional distribution of the response given a document w 2 RW: p(yjw). Due to the high dimensionality of textual data, classical regression analysis is not capable of efficiently estimating the conditional distribution, because an accurate estimation would require the sample size D to grow exponentially in the number of words W, which imposes both computational and statistical challenges. To achieve dimension reduction, IR estimates the inverse conditional distribution of the document given the response p(wjy), because this inverse conditional distribution proves to lie on a lower dimensional subspace.

**Nonparametric Topic Modeling:**

When the concentration parameter is small, the discrete distributions mostly concentrate on a few topics. As the concentration parameter increases, the discrete distributions gradually spread out probability weights to other topics. The base distribution determines the topic space and the expectation of the discrete distributions drawn from the DP. Modeling topics with the DP is advantageous because the discrete distribution drawn from the DP has a unique combination of properties. First, the topics drawn from this discrete distribution exhibit the clustering property; this property allows the words within a document to be clustered according to different topics. Second, the number of topics does not need to be specified a priori and can potentially grow with the size of corpus.

**E-Commerce Testbeeds:**

Based on the design of HDP-IR, three experiments were conducted to evaluate the predictive performance of our proposed model. The first experiment was intended to evaluate how the topic modeling component in HDP-IR improved the predictive performance. We compared our proposed model to various state-of-the-art non-topic-based models. The second experiment compared the nonparametric STMs with parametric STMs to evaluate how the nonparametric technique helped improve the predictive performance. The third experiment sought to assess how the sufficient dimension reduction generated by the inverse regression component in HDP-IR contributed to the predictive performance. To this end, we compared the HDP-IR model to various state-of-the-art supervised topic models, such as sLDA and DMR. For all experiments, we evaluted the models through five-fold cross validation to prevent the evaluation bias induced by model misspecification.

**Algorithm:**

**HDP-IR Prediction Algorithm:**

We develop an efficient inference algorithm for fitting the HDP-IR model based on the Stochastic Variational Inference (SVI) framework. The traditional variational inference algorithm needs to perform coordinate ascent over both document-level variables for all documents (i.e., E-step) and corpus-level variables in each iteration (i.e., M-step). When the traditional variational algorithm examines large corpora containing hundreds of thousands of documents, the computational complexity associated with the E-step grows significantly. Based on stochastic optimization, SVI incorporates random subsampling into the E-step and then uses the resulting accumulated document-level sufficient statistics to optimize the corpus-level variables through natural gradient ascent.

****

**Future work:**

It provides an important step for future work seeking to study the user-generated text from a topic perspective. Based on our study, we have identified a few future research directions. In this study, we focused on predicting the univariate response. To inform multi-task learning, we would like to examine whether the multivariate response also fits in HDPIR. We are also interested in exploring the role of additional factors such as the temporal factor in the extension of our proposed model. Moreover, the topic sharing idea is applicable to many other research domains, such as audio and image analysis. We therefore intend to examine the generalization of HDP-IR in such data. Furthermore, since the SDR of the HDP-IR model retains complete information about the response, we are interested in assessing whether the SDR can be used as a measurement of the document to improve the explanatory models in future social science research.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion:**

In this study, we proposed a novel nonparametric supervised topic model, HDP-IR. HDP-IR leverages the nonparametric topic modeling approach to determine the topic structure from the data. Extending the inverse regression model, HDP-IR makes predictions with sufficient dimension reduction (SDR) of the document to improve the predictive performance. Further, HDP-IR is able to examine largescale corpora containing millions of documents with the help of a novel efficient inference algorithm based on the state-of-the-art Stochastic Variational Inference. Experimental results revealed that the proposed HDP-IR model significantly outperformed existing supervised topic models. The results also suggested that both the nonparametric topic modeling component and SDR could improve the predictive performance. To the best of our knowledge, the proposed HDP-IR model is the first nonparametric topic model leveraging SDR to improve prediction accuracy.