A Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text

Yoni Halpern HALPERN@CS.NYU.EDU
New York University, New York, NY

Steven Horng SHORNG@BIDMC.HARVARD.EDU
Larry A. Nathanson LNATHANS@BIDMC.HARVARD.EDU
Nathan I. Shapiro NSHAPIRO@BIDMC.HARVARD.EDU
Beth Israel Deaconess Medical Center, Boston, MA

David Sontag DSONTAG@CS.NYU.EDU
New York University, New York, NY

Abstract

Much of clinical data is free text, which is challenging to use together with machine learning, visualization tools, and clinical decision rules. In this paper, we compare supervised and unsupervised dimensionality reduction techniques, including the recently proposed sLDA and MedLDA algorithms, on clinical texts. We evaluate each dimensionality reduction method by using them as features for two important prediction problems that arise in emergency departments: predicting whether a patient has an infection, which can progress to sepsis, and predicting the likelihood of a patient being admitted to the Intensive Care Unit (used for risk stratification). We find that, on this data, existing supervised dimensionality reduction techniques perform better than unsupervised techniques only for very low dimensional representations.

1. Introduction

Clinical text contains a lot of information about the patient and his or her context which is not captured in measured quantities like vital signs. Extracting useful information from this clinical data can be difficult because it is usually recorded as unstructured free text. Learning a low-dimensional representation of this text that captures the important information is useful for machine learning in the clinical setting with applications such as learning better clinical decision rules, information retrieval tasks and building smart surveillance and alert systems. In these settings, the learned low dimensional model can be viewed as a compact representation of the patient (or part thereof) and used as a starting point for more complex reasoning. In addition, the learned representation itself can be useful as a tool for data visualization and exploration. We focus on topic models because they are well-suited for text, and because the topics that are learned are interpretable and have face value to the clinician.

Appearing in the *ICML 2012 Workshop on Clinical Data Analysis*, Edinburgh, Scotland, UK, 2012.

In this paper we compare supervised and unsupervised dimensionality reduction techniques on a corpus of triage notes written by emergency department nurses. The triage note is available for all patients immediately upon arrival to the emergency department. Our intent is to use the learned representations in prediction algorithms early in a patients course to facilitate the clinical workflow. We test the utility of the low dimensional representations as features in two important prediction tasks that arise in emergency departments.

In the first task, we attempt to determine whether the patient is suspected to have infection and is thus at risk for developing sepsis, a severe systemic response to infection. The hospital mortality rate for severe sepsis is between 30-50%, leading to an estimated 751,000 deaths nationally (Angus et al., 2001). Early intervention has been shown to be useful in improving outcomes, so it is important to quickly identify patients with infections to prevent them from developing sepsis.

In the second task, we attempt to determine whether the patient will be admitted to the ICU from the emergency department. For this task we only consider patients who were admitted to the hospital from the emergency department. This is a more difficult prediction problem because it removes many of the easy cases, making the task closer to the real clinical setting where clinicians need to decide, based on the severity of the patient's condition, whether a patient should be admitted to the floor or to the intensive care unit.

The two tasks assess whether the representation of the triage note is useful for extracting two different types of information about the patient. The first demonstrates an understanding of the patient's complaint. The second deals with assessing the acuity of the patient's problem. These are not the only questions that would be interesting to a clinician, but we believe that a low dimensional representation of the patient that retains this information is doing a good job at distilling the important information from the note and would be a useful representation of the patient for many tasks.

We currently apply very limited preprocessing to the text, using only simple tokenization and removal of stop words. More advanced natural language processing, including negation detection and dependency parsing, would likely improve performance by providing a richer high-dimensional representation of the text. These could be used together with the dimensionality reduction methods studied in this paper.

The primary measure of effectiveness that we use is the area under the ROC curve (AUC). Due to the imbalance between the number of positive and negative cases, this ranking measure is more clinically useful than simply measuring accuracy of the classifiers.

1.1. Data Set

The triage notes come from a 55,000 visits/year Level 1 trauma center and tertiary academic teaching hospital. All consecutive ED patient visits between 12/16/2008 and 10/1/2010 were included, and no visits were excluded. A total of 94,973 patient records were randomly divided into training (60,793), validation (15,199), and test (18,981). The study was approved by our institutional review board.

For the ICU prediction task, only patients admitted to the ED were included for training and evaluation purposes. The data consists of train (24,280), validate (6,226), and test (7,547). In both tasks the class imbalance is approximately 10:1 negative to positive cases. The empirical results presented in this paper only use the train and validate sets; we withhold the test set for use in future studies.

HEADACHE Pt AOx3 presents 2 weeks of intermittent headaches, hx migraines, left sided, denies vision changes, denies weakness, denies numbness, reports intermittent tingling in left arm, none now, speech clear, no facial droop, c/o nausea, no vomitting
LEFT LEG PAIN 30 yr old woman from home afet ra trip and fall down 5 steps, no loc, no head strike. Unable to weight bear. Pain in left leg from knee to ankle, some pain also in left knee. No obvious deformity.
FOOT INFECTION Pt here from ----- hosp.with ?osteomyelitis. Footis pink swollen and warm to the touch on the right foot. Denies fevers at home. hx of multiple infections after a mvc ankle break in ...
SORE THROAT pt c/o sore throat x 1 week - pt states took antibiotic x 5 days after initiation of sore throat and sx resolved and now back after completed antibiotics - pt tolerating po fluids yet c/o pain on swallowing
CHEST PAIN presents with left sided chest pain intermittant described as gas pain today pain reoccured during episode of stress, developed fluttering in left chest with left arm pain. Denies n/v/d or dyspnea

Figure 1. Examples of triage notes from the dataset. The capitalized words are the chief complaints. The notes tend to be concisely written with multiple abbreviations and misspellings. Identifying information has been removed.

The ground truth labeling for whether a patient was suspected to have an infection is determined by checking whether any one of their ED ICD-9-CM discharge diagnoses was in the Angus sepsis abstraction criteria, a list of ICD-9-CM codes that correspond to diagnoses consistent with infection (Angus et al., 2001). Although noisy, ICD-9-CM discharge codes are the best routinely captured data that reveals whether an infection was suspected. The sensitivity of the ICD-9-CM codes, capturing even suspected infection, is useful in this setting because early interventions to prevent the progression of infection are administered to patients where infection is suspected. In addition, current work in quality improvement advocates evaluating clinical pathways based on suspicion of disease, rather than ultimate diagnosis.

Triage notes tend to be concisely written (10-30 words) with many variants of shortforms and misspellings. Figure 1 shows some examples of triage notes. Each note contains a nurse's free text description of a patient at the time of arrival in the emergency department.

1.2. Clinical Applications

The early introduction of predictive algorithms into clinical practice will provide critical insights to guide their development. We have piloted the real-world clinical applicability of these dimensionality reduction methods with two preliminary tests.

In Figure 2, we provide a rank order list of patients by probability of ICU admission, which we use as a surrogate measure for acuity. First, this allows us to prospectively validate the accuracy of our algorithms against the clinical judgment of clinicians currently taking care of those patients. High acuity patients

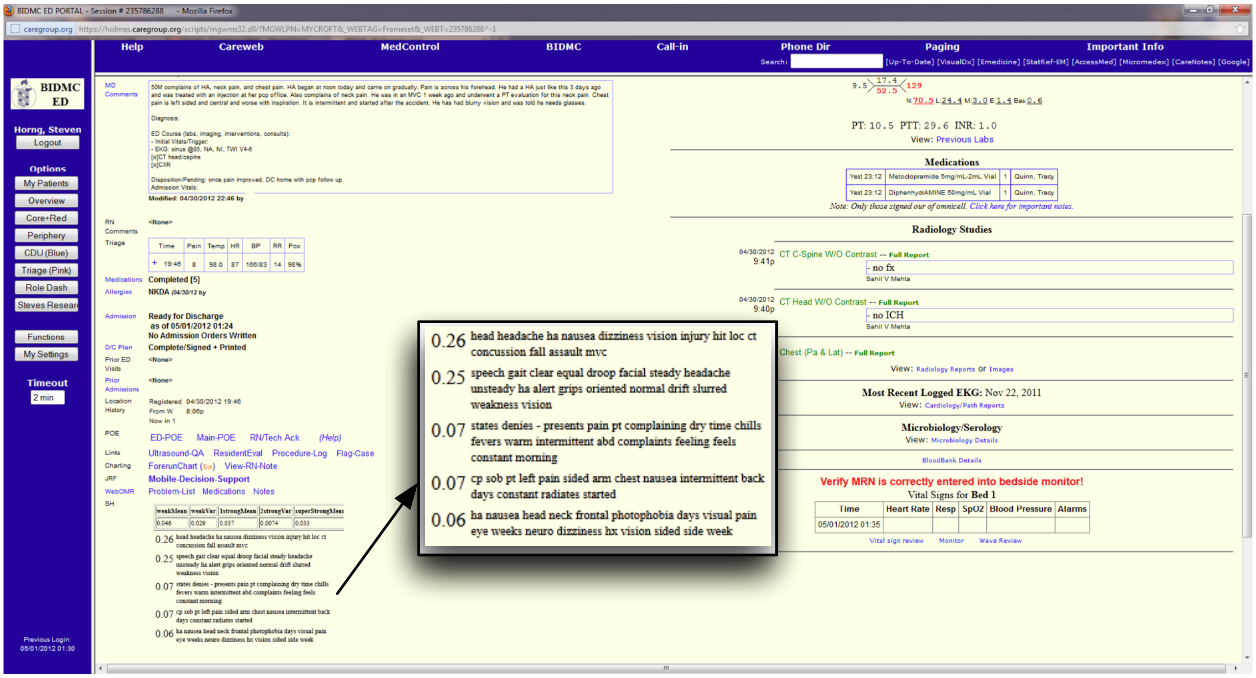


Figure 3. Screenshot of detailed patient view for the informatics system used in the emergency department. During our pilot, for each patient we display the most probable topics, computed in real-time using the available clinical text. This allows us to, during algorithm development, quickly assess the ability of the dimensionality reduction methods to provide interpretable and accurate results.

that are ranked low (and the reverse) can quickly be discerned, and a deeper analysis can then be initiated to determine the cause of this erroneous prediction. Practically, a rank order listing of patients by acuity is continuously being calculated by the emergency physician. However, this task is dependent on the abilities of the individual clinician, and is prone to systemic errors. Such a system is particularly vulnerable during shift changes when the entire personnel of an emergency department turns over. The new clinician must quickly determine this rank order listing without having personally examined many of these patients. Our rank order listing using our algorithms formalizes this often subconscious task, ensuring that higher acuity patients are not overlooked.

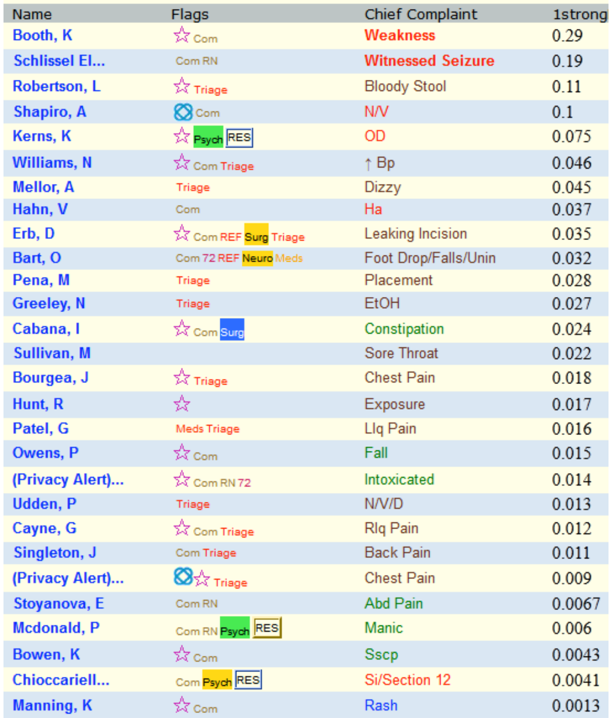
In Figure 3, we show how output from machine learning algorithms can be integrated into existing patient information displays. These non-invasive methods allow clinicians to evaluate the output of these algorithms in the context of patient care. Real-time evaluation is important as it uses all available data, not just data that is captured. Furthermore, it is well documented that clinicians have different behaviors, attitudes, and information needs when they are working clinically than when they are not. Even the best clinical simulation can not reproduce the unscheduled interruptions, emotions, and stress that is inherent in clinical care. Practically, this would be the same location where output from these machine learning algorithms can be used to trigger alerts, reminders, and decision support.

2. Related Work

In this section we describe related work on dimensionality reduction applied to clinical text, in addition to related work on the particular tasks of predicting whether a patient has an infection and whether they will be admitted to the ICU after arriving in the emergency department.

Recent work has looked at applying topic models to clinical text. Salleb-Aouissi et al. (2011) use latent Dirichlet Allocation (LDA) as an exploration tool for better understanding infant colic from pediatric notes. Perotte et al. (2011) propose a new supervised learning algorithm for LDA for hierarchically-structured prediction tasks. They apply their algorithm to the task of automatically assigning ICD-9 codes (used for billing) from discharge summaries. Their probabilistic model reduces to sLDA (Blei & McAuliffe, 2007) when applied without the hierarchy.

In contrast to these earlier works, the clinical text that we consider (triage notes) has substantially shorter



Name	Flags	Chief Complaint	1strong
Booth, K	Com	Weakness	0.29
Schlissel El...	Com RN	Witnessed Seizure	0.19
Robertson, L	Triage	Bloody Stool	0.11
Shapiro, A	Com	N/V	0.1
Kerns, K	Psych RES	OD	0.075
Williams, N	Com Triage	↑ Bp	0.046
Mellor, A	Triage	Dizzy	0.045
Hahn, V	Com	Ha	0.037
Erb, D	Com REF Surg Triage	Leaking Incision	0.035
Bart, O	Com 72 REF Neuro Meds	Foot Drop/Falls/Unin	0.032
Pena, M	Triage	Placement	0.028
Greeley, N	Triage	EtOH	0.027
Cabana, I	Com Surg	Constipation	0.024
Sullivan, M		Sore Throat	0.022
Bourgea, J	Triage	Chest Pain	0.018
Hunt, R		Exposure	0.017
Patel, G	Meds Triage	Llq Pain	0.016
Owens, P	Com	Fall	0.015
(Privacy Alert)...	Com RN 72	Intoxicated	0.014
Udden, P	Triage	N/V/D	0.013
Cayne, G	Com Triage	Rlq Pain	0.012
Singleton, J	Com Triage	Back Pain	0.011
(Privacy Alert)...	Triage	Chest Pain	0.009
Stoyanova, E	Com RN	Abd Pain	0.0067
Mcdonald, P	Com RN Psych RES	Manic	0.006
Bowen, K	Com	Sscp	0.0043
Chioccariell...	Com Psych RES	Si/Section 12	0.0041
Manning, K	Com	Rash	0.0013

Figure 2. De-identified screenshot from our pilot of a risk stratification algorithm in the emergency department, which uses the dimensionality reduction methods described in this paper (all names are fake). Patients are ordered by probability of ICU admission (shown on the far right column), which we use as a surrogate measure for acuity.

documents. Triage notes are an important source of information since they are available at the very beginning of a patient's hospital visit. This emphasis on information available very early in the patient's visit constitutes a significant departure from the standard methods of risk stratification and early detection of sepsis which focus on more objective data sources such as white blood cell count and lactate, available later in the patient's course. For our applications in facilitating clinical workflow, it is critical to use triage time information, since diagnostic and treatment plans are often already formulated by the time other sources of information become available.

To our knowledge, our work is the first to make substantial use of triage nursing notes in the emergency department to predict whether a patient has a suspected infection. Several authors have considered the problem of early identification of sepsis in the emergency department (e.g., Shapiro et al., 2003; Nelson et al., 2011). However, prior work has primarily focused on simple hand-written rules that are unable to take advantage of the unstructured clinical text.

We are also the first, to our knowledge, to make use of unstructured clinical text for risk stratification in the emergency department. Earlier work focused primarily on using vital signs alone, or together with the chief complaint (2–3 words).

Although this paper focuses on comparison of supervised and unsupervised dimensionality reduction techniques, we emphasize that *all* of these methods achieve state-of-the-art performance on *both* tasks compared to classifiers trained on other information available at triage time that ignore the unstructured text.

3. Unsupervised Methods

3.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (Blei et al., 2003) describes a generative model for the creation of a textual corpus. Formally, the generative process for a single triage document is as follows: We start by specifying a number of topics, T. A distribution θ over the T topics is sampled from a Dirichlet prior with hyperparameters α. This distribution specifies, roughly, the fraction of words in the triage note that will correspond to each topic. Next, for each word i, we choose a topic z_i according to the multinomial distribution parameterized by θ. Finally, we sample a word w_i from a multinomial distribution parameterized by β_{z_i}.

The learning task is to find a latent representation, parametrized by values for the hyperparameters α and β to maximize the likelihood of the data under the model. Once the hyperparameters α and β are learned, inference can be performed to compute a posterior θ vector for a new document which can be interpreted as a low dimensional summarization of the document. Exact inference in the LDA model and its generalizations is NP-hard, so approximate inference algorithms are used (Sontag & Roy, 2011).

3.2. Singular Value Decomposition (SVD)

Singular value decomposition can be used to find a set of orthogonal basis vectors such that the projection of the word-document counts matrix onto that basis captures the maximum variance of the data. We present it here as a baseline for the accuracy of prediction after dimensionality reduction. In contrast to the LDA models, we were unable to find a reasonable way to visualize the latent dimensions learned by SVD.

4. Supervised Methods

Unsupervised learning attempts to find a latent structure that best explains the text data, either in terms of likelihood (as in LDA) or ℓ_2 error (as in SVD). Supervised variants of LDA attempt to use labeled data to learn a latent description of the data that is useful for a particular task.

4.1. sLDA

sLDA (Blei & McAuliffe, 2007) adds the supervised signal into the standard generative model of LDA so that for each document d, a response y_d is sampled conditional on the topic assignments $z_{1:N}$. For binary classification, this distribution is given by $\Pr(y_d = 1 \mid z_{1:N}) = \frac{1}{1+\exp(-\eta^T \bar{z})}$, where $\bar{z}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[z_i = k]$ with k indexing the topics. η denotes a weight vector to be learned.

Similar to LDA, the learning task involves finding the hyperparameters α, β and η that maximize the likelihood of the data under the model.

4.2. MedLDA

Instead of having a fully generative model, MedLDA (Zhu et al., 2009) optimizes an objective that balances the discriminative ability of the model with the likelihood of the data under the LDA model. Learning in MedLDA corresponds to maximizing the objective $L(q;\alpha,\beta) - C\sum_{d=1}^{D}\xi_d - ||\eta||^2$, where $L(q;\alpha,\beta)$ is a variational lower bound on the likelihood of the data, and ξ_d is a measure of the prediction error on document d. For each d, we have the constraints $y_d\,(\eta^T \mathbb{E}_q[\bar{z}_d]) \geq 1 - \xi_d$, and $\xi_d \geq 0$. As in the other models, the goal of learning is to find values for the hyperparameters α, β and η. The weighting between the two objectives, C, is chosen on a validate set.

5. Implementation

We computed the singular value decompositions using the *irbla* package for R (Baglama & Reichel, 2011). We apply SVD to the term-document counts matrix after normalizing so that the counts for each document sum to one. We also tried using inverse document frequency weighting (tf-idf), but found that it gave no improvement. We used a linear classifier, trained with a support vector machine (SVM), to do prediction using the unsupervised learning algorithms. SVM training was done with SVM^{perf} (Joachims, 2005), optimizing a lower bound on the AUC rather than accuracy. The regularization constant was chosen by trying values of 10^k for $k \in \{-4,-3,-2,...,3,4\}$.

The sLDA models were learned using the *lda* package for R (Chang, 2011). We modified the code to do non-uniform hyperparameter optimization for α. In the experiments, we used 3 EM steps and 100 Gibbs iterations per E step. We tried a range of values up to 1000 Gibbs iterations, and no improvement was observed. 3 EM steps were chosen because we noticed that any further EM steps caused the AUC on validate data to either decline or stay almost constant.

The MedLDA model was learned using the authors' provided code. We modified the code to use a re-scaled loss function to account for class imbalance. The C parameter was chosen using the validate set by trying values of 10^k for $k \in \{-1,0,...,3\}$. Similar to Zhu et al. (2009), we do a second step where we re-learn the weight vector using a SVM (in our case, optimizing AUC) on the low dimensional representation of the training data output by MedLDA.

The Gibbs sampling LDA model was learned with Mallet (McCallum, 2002). The mean-field LDA model was learned using the MedLDA code with the supervised term removed. This is equivalent to the variational EM algorithm for LDA, with non-uniform hyperparameter optimization for the Dirichlet.

6. Results

We consider the unsupervised methods SVD and LDA to be baseline algorithms, and test whether supervised methods improve performance. We also provide the performance of a bag-of-words model which represents a document as a feature vector, with one dimension per word whose value is the unnormalized term count. The bag-of-words model is extremely high dimensional; we do not intend to compare it to the other models directly, but rather to provide a measure of how much information is lost in the dimensionality reduction.

Since there are substantial performance differences between implementations that use mean-field and sampling-based approximations for inference, we only compare methods that use the same type of approximate inference scheme.

The results are presented in Figures 4 and 5. Bag-of-words obtains better AUC than all of the low-dimensional representations for both prediction tasks. We emphasize that the low-dimensional representations still have several advantages, such as providing a latent representation of the patient for use in other applications (e.g., automated patient surveillance algorithms) and for visualization. As the number of topics increases we expect this gap to close, and indeed observe this with respect to the infection prediction task. Interestingly, for the ICU prediction task, even after 100 topics there is still a substantial gap. This deserves further study, and may motivate completely new latent variable learning algorithms.

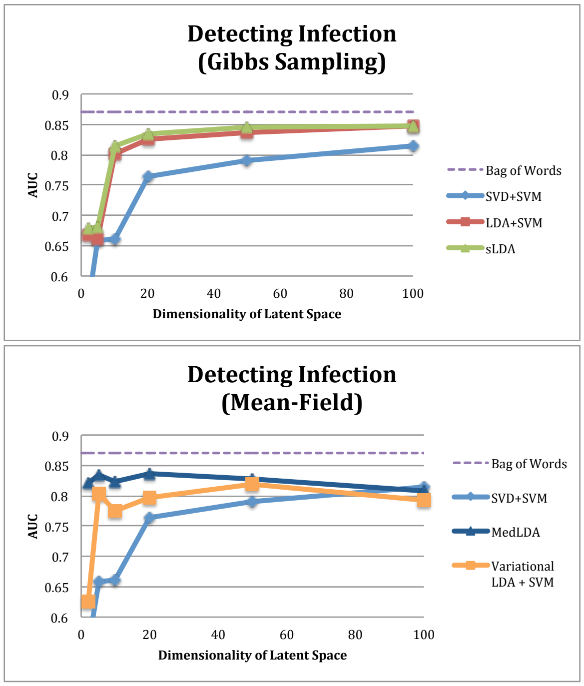


Figure 4. Comparison of AUC results on validate data for detecting infection. The ordering between LDA+SVM and sLDA should not be interpreted as significant due to slight differences in the software packages.

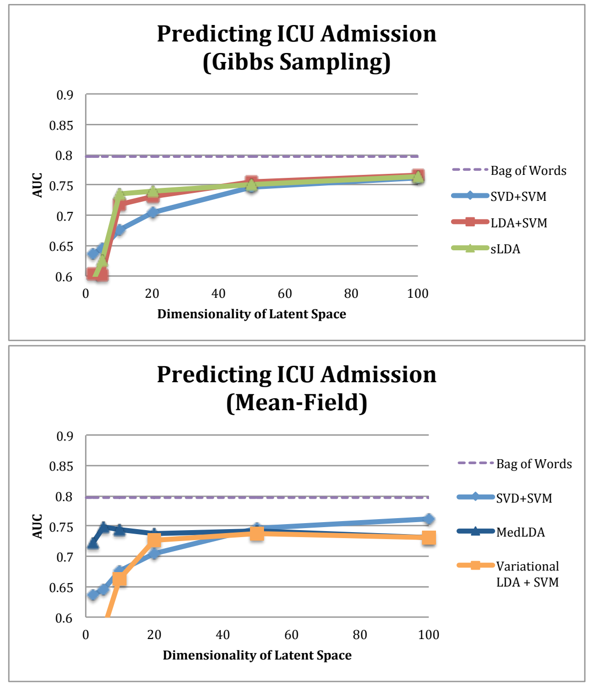


Figure 5. Comparison of AUC results on validate data for predicting whether a patient will be sent to the ICU, given that they are admitted to the hospital.

As we mentioned earlier, these results are given on the validate data. This is a slightly optimistic performance measure, since we also selected several model parameters using the same data. We computed the test error in several cases and observed that the test error and the validate error are very close.

7. Challenges of Supervised Dimensionality Reduction

We had expected the supervised methods of dimensionality reduction to outperform the unsupervised methods by finding a representation that better conserves the information important for the task at hand. While MedLDA outperforms other methods when using a small number of topics, once more than 50 topics are used, the supervised signal stops being useful. When a small number of topics are used, it is extremely important that the model choose the right topic representations in order to capture information about the signal that we are interested in. We illustrate this in Table 1 by showing three of the 10-topic models. As the number of topics increases, any reasonable description of the data that uses a large number of fine grained topics works well for our tasks.

We found that sLDA performed slightly worse than the unsupervised Gibbs sampling-based LDA. A possible cause for this is that sLDA jointly maximizes the likelihood of the data, treating the response variable in essence as another word of the document. In practice, the contribution to the likelihood of the words can completely outweigh the contribution from the responses. In our experiments we found that the models found by sLDA were very similar to each other, even when using different supervised signals for the different tasks. They were also similar to the models found by unsupervised LDA.

MedLDA addresses the weighting issue by introducing a free parameter that weighs the discriminative performance of the model against the likelihood of the data under the model, as computed using a mean-field variational approximation. This can introduce an interesting type of overfitting. When the discriminative term of the objective is weighted strongly, it biases the variational posterior distribution toward one that would minimize the hinge loss, but it can do so at the expense of increasing the divergence between the variational distribution and the true distribution. This succeeds at finding a low-dimensional representation of the training data that is (nearly) separable during training time, i.e. where $\sum_d \xi_d$ is very small. However, this is not the same low-dimensional representation that is found at test time when the supervised

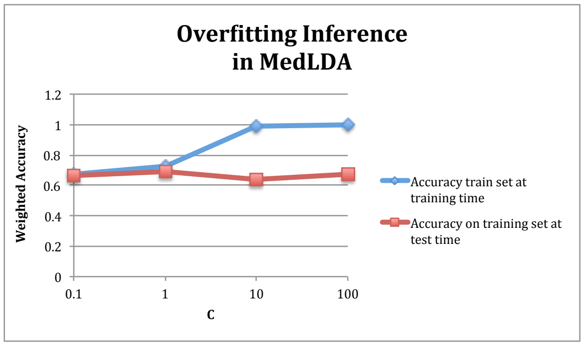


Figure 6. Comparison of the weighted accuracy on the training set during training and when the same data is presented at test time without labels. The accuracy is calculated with positive examples weighted ten times more than negative examples. The x-axis shows the C parameter that balances between the discriminative term and the likelihood of the data. When C is large, the accuracy on the train set during training is near perfect, but is much lower when the same data is presented at test time. Since the variational inference procedure during training is influenced by the supervised signal, it can find distributions that are far from the true distribution, and which are not recovered at test time.

signal is removed, leading to poor performance even on the training set (see Figure 6).

8. Conclusion

We compared supervised and unsupervised dimensionality reduction techniques, including the recently proposed sLDA and MedLDA algorithms, on triage nursing notes. For very low-dimensional representations, we find that the supervised MedLDA approach outperforms unsupervised approaches. However, we find that supervised dimensionality reduction techniques perform no better than unsupervised approaches when using a representation with more than 50 dimensions. For many practical applications involving unstructured clinical text we believe a latent representation of 50 dimensions or more is reasonable, and unsupervised LDA is the simplest and most effective representation.

Starting with a richer, high dimensional, representation built with advanced NLP tools such as negation detection would likely improve the performance of all of the low dimensional models, and is a promising direction for future work. Finally, the large gap between the bag-of-words model and all of our low dimensional representations in the ICU task suggests that new methods are needed to learn low dimensional latent variable models that can describe the patient well.

Acknowledgments

This work was supported in part by CIMIT Award No. 12-1262 under U.S. Army Medical Research Acquisition Activity Cooperative Agreement W81XWH-09-2-0001, and by grant UL1 TR000038 from the National Center for the Advancement of Translational Science (NCATS), National Institutes of Health. The information contained herein does not necessarily reflect the position or policy of the Government, and no official endorsement should be inferred.

References

Angus, D., Linde-Zwirble, W., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. Epidemiology of Severe Sepsis in the United States: Analysis of Incidence, Outcome, and Associated Costs of Care. *Critical Care Medicine*, 29(7):1303–10, July 2001.

Baglama, Jim and Reichel, Lothar. *irlba: Fast partial SVD by implicitly-restarted Lanczos bidiagonalization*, 2011. R package version 1.0.1.

Blei, David M. and McAuliffe, Jon D. Supervised topic models. In *NIPS*, 2007.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Chang, J. *lda: Collapsed Gibbs sampling methods for topic models.*, 2011. R package version 1.3.1.

Joachims, T. A Support Vector Method for Multivariate Performance Measures. In *ICML*, pp. 377–384. ACM Press, 2005.

McCallum, A. MALLET: A Machine Learning for Language Toolkit. 2002.

Nelson, J., Smith, B., Jared, J., and Younger, J. Prospective Trial of Real-Time Electronic Surveillance to Expedite Early Care of Severe Sepsis. *Annals of Emerg. Medicine*, 57(5):500–4, May 2011.

Perotte, A., Bartlett, N., Elhadad, N., and Wood, F. Hierarchically Supervised Latent Dirichlet Allocation. *Neural Information Processing Systems*, 2011.

Salleb-Aouissi, A., Radeva, A., Passonneau, R., Xie, B., Khattak, F., Ashish Tomar, H., Waltz, D., McCord, M., McGurk, H., and Elhadad, N. Diving into a large corpus of pediatric notes. *ICML Workshop – Learning from Unstructured Clinical Text*, 2011.

Shapiro, N.I., Wolfe, R., Moore, R., Smith, E., Burdick, E., and Bates, D. Mortality in emergency department sepsis (MEDS) score: a prospectively derived and validated clinical prediction rule. *Critical Care Medicine*, 31(3): 670–675, 2003.

Sontag, D. and Roy, D. Complexity of inference in latent dirichlet allocation. In *Neural Information Processing Systems 24*, pp. 1008–1016. 2011.

Zhu, J., Ahmed, A., and Xing, E. MedLDA: maximum margin supervised topic models for regression and classification. In *ICML*, pp. 1257–1264, 2009.

Unsupervised LDA **MedLDA (infection)** **MedLDA (ICU)**

Table 1. Illustration of three different 10-topic LDA models learned using triage notes. The size of a word is proportional to its probability. The left column's topics were learned by unsupervised LDA (variational algorithm) over all patient notes. The middle column's topics were learned by MedLDA over all patient notes with the response variable being infection (topics are sorted by the linear SVM weight vector, with the largest weights on the bottom). The right column's topics were learned by MedLDA over admitted patient notes with the response variable being ICU admission (also sorted). Many of the topics with the large weights in each of the supervised models do not have clear analogues in the other models, suggesting that the supervised model has learned a representation specifically suited to its task.