15

Mixed Membership Classification for Documents with Hierarchically Structured Labels

Frank Wood

Department of Engineering, University of Oxford, Oxford, OX1 3PJ, UK

Adler Perotte

Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA

CONTENTS

15.1 Introduction 306
15.1.1 Background 306
Latent Dirichlet Allocation 306
Supervised Latent Dirichlet Allocation 307
15.2 Hierarchical Supervised Latent Dirichlet Allocation 307
15.2.1 Generative Model 308
15.2.2 Dealing with Label Imbalance 310
15.2.3 Intuition 310
15.3 Inference 311
15.3.1 Gibbs Sampler 311
15.4 Example Applications 312
15.4.1 Hospital Discharge Summaries and ICD-9 Codes 312
15.4.2 Product Descriptions and Catalogs 314
15.4.3 Comparison Models 315
15.4.4 Evaluation and Results 316
15.5 Related Work 318
15.6 Discussion 319
Appendix 320
References 321

Placing documents within a hierarchical structure is a common task and can be viewed as a multi-label classification with hierarchical structure in the label space. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. We present a model for hierarchically and multiply labeled bag-of-words data called hierarchically supervised latent Dirichlet allocation (HSLDA). Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-words data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

15.1 Introduction

Documents frequently come with additional information like labels or popularity ratings. Contemporary examples include the product ratings that accompany product descriptions, the number of “likes” that webpages have attracted, grades associated with assigned essays, and so forth.

This chapter covers one way to jointly model documents and the labels applied to them. In particular we focus on our own work on modeling documents with more complicated labels that themselves possess some kind of structural organization. Consider typical product catalogs. They usually contain text descriptions of products that have been organized into hierarchical product directories. The situation of a product into such a hierarchy (the path or paths in the product hierarchy that lead to it) can be thought of as a structured label. Jointly modeling the document and such a label is useful for automatically labeling new documents (corresponding in this example to automatically situating a new product in the product directory) and more.

Collections of hierarchically labeled documents abound, text and otherwise. We will consider hierarchically labeled patient clinical records in later sections. Applications like situating web-pages in hierarchical link directories are left for others to explore.

Because text documents are notoriously difficult to directly model we take an approach common to other chapters in this book. We use a mixed membership model of the text document to represent the document as a bag-of-words drawn from a document-specific mixture of topic distributions. The modeling choices we make in relating this representation to structured labels follows, as does its relationship to prior art.

15.1.1 Background

Mixed membership models, including the model upon which we build, latent Dirichlet allocation (LDA) (Blei et al., 2003), have been reviewed in other chapters. The key property we exploit for purposes of classification is that LDA provides a way to extract a latent, low-dimensional representation of text and other documents consisting of the frequency of word assignments to the topics that are assumed to have generated them. A topic is a distribution over words. Each document is a bag-of-words drawn from a document-specific mixture of topics.

Building a joint model of documents and labels using this representation is not new. It was first introduced by Blei and McAuliffe in a paper on “supervised” latent Dirichlet allocation (SLDA) (Blei and McAuliffe, 2008). SLDA built on LDA by incorporating “supervision” in the form of an observed exponential family response variable per document.

Latent Dirichlet Allocation

To explain both SLDA and to set the stage for our work, it helps to introduce our notation for LDA. Assume that there are K topics. Let $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$ be “topic” k, i.e., a distribution over a finite set of words. Here, V simply labels the variables to indicate that they have to do with the vocabulary. The vector $\mathbf{1}_V$ consists of all ones and has length equal to the size of the vocabulary. The distribution Dir is the Dirichlet distribution. The constant γ controls the smoothness of the inferred topics. Larger values lead to smoother topic estimates. Let $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K\left(\alpha' \mathbf{1}_K\right)$ be a “global” distribution over topics where K indicates that the distribution and variable sizes are equal to the number of topics K and α' controls the relative proportion of topics globally, large α' leading to all topics being roughly responsible for the same number of words. Intuitively, $\boldsymbol{\beta}$ is something like the average topic proportion independent of any particular document. Per document d, topic distributions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K\left(\alpha \boldsymbol{\beta}\right)$ are modeled as being deviations from the global distribution over topics where larger values of α result in all documents' topic distributions being more similar. We will use $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ to indicate which topic generated the nth word of document d. Drawing the nth word in document d from the indicated topic $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$ completes our notation of standard LDA.

Supervised Latent Dirichlet Allocation

Supervised LDA adds another per document observation, a label y_d. It is modeled as being generated by a generalized linear model (GLM) $y_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}, \delta \sim \text{GLM}(\bar{\mathbf{z}}_d, \boldsymbol{\eta}, \delta)$. Brushing aside exponential family and GLM link function generalities, what SLDA does is regress labels against the empirical distributions of assigned topic indicator variables $\bar{\mathbf{z}}_d = \{\bar{z}_{1,d}, \ldots, \bar{z}_{K,d}\}$, where $\bar{z}_{k,d} = \frac{\sum_n \mathbb{I}(z_{n,d}=k)}{\sum_n \sum_j \mathbb{I}(z_{n,d}=j)}$ is the fraction of words assigned to topic k and $\mathbb{I}(\cdot)$ is the indicator function that returns one if its argument is true. If the document labels are real-valued then one example choice for the regression relationship would be $y_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta} \sim \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}, \delta)$. Using a generalized linear model in the exponential family to parameterize this regression relationship allows for a wide variety of distributions over different kinds of label spaces to be represented in the same mathematical formalism. A variational expectation maximization algorithm was proposed in Blei and McAuliffe (2008) to learn model parameters. Experimental results in Blei and McAuliffe (2008) showed both excellent out-of-sample label prediction and improved topics. Topic improvement was measured by using the empirical topic proportions from SLDA as features for external, discriminative approaches to label prediction. Regression models built on SLDA topics outperformed the same built on LDA derived topics.

In one sense, SLDA is more general than the model we present in this chapter, namely, the labels need not be categorically valued. The main subject of this chapter, a generalization of SLDA called hierarchically supervised LDA (HSLDA) (Perotte et al., 2011) does not deal with real-valued labels, however, it is more general than SLDA in the case of categorical labels. The exponential family/GLM regression framework can theoretically account for multivariate labels and potentially even structured categorical labels. HSLDA is, however, a specific, practical way to model with structured categorical labels. Because we focus on hierarchically structured categorical labels, we refer to our model as a mixed membership hierarchical classification model.

15.2 Hierarchical Supervised Latent Dirichlet Allocation

This model (HSLDA) is designed to fit hierarchically, multiply-labeled, bag-of-word data. We call groups of bag-of-words data documents (unordered words in text documents, bag of visual feature representations of images, etc.). Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$.

Let the set of labels be $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. A label in HSLDA corresponds to a node in the graphical model in Figure 15.1. A label can either be observed or unobserved. Documents can be multiply labeled, meaning that subsets of label nodes in the graphical model in Figure 15.1 can have observed values. Each label $l \in \mathcal{L}$, except the root, has a parent $\text{pa}(l) \in \mathcal{L}$ also in the set of labels. We will, for exposition purposes, assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Readers may wish to consult Figure 15.2 or Figure 15.3, each of which is a label-tree graphical model.

In a label forest a node may be observed (the label was "applied" and, for instance, was observed to have value 1), unobserved and unknown, or unobserved and constrained to be either -1 or 1 by the structure of the label space. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label applies to document d or not. In most cases $y_{l,d}$ will be unobserved; in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we demonstrate, only positive labels are observed. This may not be true of all applications, however, positive-only label imbalance is a common problem. How we solve this problem will be discussed later.

The constraints imposed by an is-a label hierarchy are that if the lth label is applied to document d, i.e., $y_{l,d} = 1$, then all labels in the label hierarchy up to the root are also applied to document d, i.e., $y_{\mathrm{pa}(l),d} = 1, y_{\mathrm{pa}(\mathrm{pa}(l)),d} = 1, \ldots, y_{r,d} = 1$. Conversely, if a label l' is marked as not applying to a document (i.e., $y_{l',d} = -1$) then no descendant label of that label can take value 1. We assume that at least one label is applied to every document. This is illustrated in Figure 15.1 where the root label is always applied but only some of the descendant labelings are observed as having been applied (diagonal hashing indicates that potentially some of the plated variables are observed).

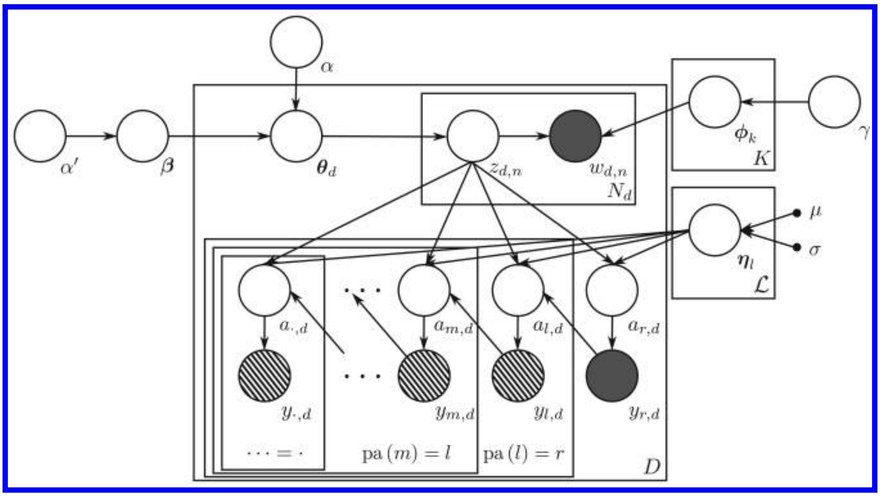


FIGURE 15.1
Hierarchically supervised latent Dirichlet allocation (HSLDA) graphical model.

In HSLDA, the bag-of-word document data is modeled using LDA with full, hierarchical topic estimation (i.e., global topic proportions are also estimated). Label responses are modeled using a conditional hierarchy of probit regressors and will be discussed next. The full HSLDA graphical model is given in Figure 15.1.

15.2.1 Generative Model

In the following box the HSLDA generative model is given for the "is-a hierarchy" set of label constraints. In the box and what follows in this chapter, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise.

HSLDA Generative Model

1. For each topic $k = 1, \dots, K$
   - Draw a distribution over words $\boldsymbol{\phi}_k \sim \mathrm{Dir}_V(\gamma \mathbf{1}_V)$.
2. For each label $l \in \mathcal{L}$
   - Draw a label weight vector $\boldsymbol{\eta}_l \mid \mu, \sigma \sim \mathcal{N}_K(\mu \mathbf{1}_K, \sigma \mathbf{I}_K)$.
3. Draw the global topic proportions $\boldsymbol{\beta} \mid \alpha' \sim \mathrm{Dir}_K\,(\alpha' \mathbf{1}_K)$.
4. For each document $d = 1, \dots, D$
   - Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K\,(\alpha \boldsymbol{\beta})$.
   - For $n = 1, \dots, N_d$
     - Draw topic assignment $z_{n,d} \mid \boldsymbol{\theta}_d \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$.
     - Draw word $w_{n,d} \mid z_{n,d}, \boldsymbol{\phi}_{1:K} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{n,d}})$.
   - Set $y_{r,d} = 1$.
   - For each label l in a breadth first traversal of $\mathcal{L}$ starting at the children of root r
     - Draw

$$a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1. \end{cases} \tag{15.1}$$

     - Apply label l to document d according to $a_{l,d}$

$$y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise.} \end{cases} \tag{15.2}$$

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \dots, \bar{z}_k, \dots, \bar{z}_K]$ is the empirical topic distribution for document d, in which each entry is the percentage of the words in that document that come from topic k, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$. As in Blei and McAuliffe (2008), the response variables are directly dependent on $\bar{\mathbf{z}}_d^T$ because this directly couples the topic assignments used to explain the words and the topic assignments used to explain the responses.

The second half of Step 4 is what is referred to as supervision in the supervised LDA literature. This is where the hierarchical classification of the bag-of-words data takes place and the is-a label constraints are enforced. For every label $l \in \mathcal{L}$, both the empirical topic distribution for document d and whether or not its parent label was applied (i.e., $\mathbb{I}(y_{\mathrm{pa}(l),d} = 1)$) are used to determine whether or not label l is to be applied to document d as well. Equations (15.1) and (15.2) comprise a probit regression model in an auxiliary variable formulation (see Appendix). Note that in the case that the parent label is applied, i.e., $y_{\mathrm{pa}(l),d} = 1$, the child label $y_{l,d}$ is applied with probability $P(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l > 0)$. This is a conditional probit regression model for classification where $\boldsymbol{\eta}_l$ are the class-conditional regression parameters. The auxiliary variables $a_{l,d}$ make inference tractable but are not fundamental to the model—only the labels and regression parameters are actually of interest.

Note that $y_{l,d}$ can only be applied to document d (set to 1) if its parent label $\mathrm{pa}(l)$ is also applied (these expressions are specific to is-a constraints but could be modified to accommodate different constraints between labels). Note that multiple labels can be applied to the same document. The regression coefficients $\boldsymbol{\eta}_l$ which generate the labels are independent a priori, however, the hierarchical coupling in this model and conditional label dependency structure induces a posteriori dependence. The net effect of this conditional hierarchy of profit regressors is that child label predictors deeper in the label hierarchy are able to focus on finding features that distinguish label paths in the tree, conditioned on the fact that all the children of any particular node are by design members of some more general parent set. One can restrict this hierarchy to a depth of one hierarchy, recovering SLDA with probit link and univariate categorical labels. Also, one can nearly as easily make the conditional classification at each node multi-class rather than single-class if more than one label at each node is required. In many cases, however, a binary indicator along with a deeper or more complex tree is sufficient.

Note that the choice of variables $a_{l,d}$ and how they are distributed were driven at least in part by posterior inference efficiency considerations (see Appendix). In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (15.5) and the regression coefficients (15.4), which are analytic. This simplifies posterior inference substantially. A review of probit regression can be found near the end of this chapter in the Appendix.

15.2.2 Dealing with Label Imbalance

In the common case where no negative labels are observed (like the example applications we consider in Section 15.4), the model must be explicitly biased towards generating negative labels in order to keep it from learning to only assign positive labels to all documents. This is a common problem in modeling with unbalanced labels. To see how this model can achieve this we draw the reader's attention to the μ parameter and, to a lesser extent, the σ parameter. Because $\bar{\mathbf{z}}_d$ is always positive, setting μ to a negative value results in a bias towards negative labelings, i.e., for large negative values of μ, all labels become a priori more likely to be negative ($y_{l,d} = -1$). We explore the effect of μ on out-of-sample label prediction performance in Section 15.4. In a very real way, μ is a knob that can be adjusted both before inference to induce a broad array of out-of-sample performance characteristics that vary along classical axes like specificity, recall, and accuracy. A similar but less principled solution can be effected by changing the decision boundary from 0 in (15.1) and (15.2). This technique can be used to vary out-of-sample label bias after learning.

15.2.3 Intuition

To help ground this abstract graphical model, recall the retail data example application. We asserted that retailers often have both a browseable product hierarchy and free-text descriptions for all products they sell. The situation of each product in a product hierarchy (often multiply situated) constitutes a multiple, hierarchical labeling $\mathbf{y}_d$ of the free-text product descriptions $\mathbf{w}_d$ for all products d. Note that a single product can be placed in the hierarchy in multiple places. This corresponds to multiple paths in the label hierarchy having labels that are all applied. HSLDA assumes that the free-text descriptions of all of the products in a particular node in the product hierarchy must be related. It also assumes that products deeper in the product hierarchy are described using language that is similar to that used to describe products in their parent classes. For instance, basketballs are probably described using language that is similar to that used to describe other basketballs, other balls, and more general sporting goods. In both the lay and technical senses, similar products should have product descriptions that share topics. If topic proportions are indicative of the text describing products that are grouped together, the key HSLDA assumption is that it should then be possible to use those proportions to decide (via probit classification) whether or not a particular product should be situated at a particular node in the product hierarchy. Conversely, that certain groups of products are known to be clustered together should inform the kinds of topics that are inferred from the product descriptions.

15.3 Inference

Our inference goal is to obtain a representation of the posterior distribution of the latent variables in the model. This posterior distribution can then be used for predictive inference of labels for held-out documents, among other things. Unfortunately, the posterior distribution we seek does not have a simple analytic form from which exact samples can be drawn. This is usually the case for posterior distributions of non-trivial probabilistic models and suggests approximating the posterior distribution by sampling.

In this section we derive the conditional distributions required to sample from the HSLDA posterior distribution using Markov chain Monte Carlo. The HSLDA sampler, like the collapsed Gibbs samplers for LDA (Griffiths and Steyvers, 2004), is itself a collapsed Gibbs sampler in which all of the latent variables that can be analytically marginalized are. Among others, the topic distributions $\boldsymbol{\phi}_{1:K}$ and document-specific topic assignment distributions $\boldsymbol{\theta}_{1:D}$ are analytically marginalized prior to deriving the following conditional distributions for sampling.

It will usually be the case that values $y_{l,d}$ will not be known for all labels $l \in \mathcal{L}$ in the space of possible labels. Values for $y_{l,d}$ that are enforced by label constraints and observed labels are set to their constrained values prior to inference and treated as observed. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed (or observed via filling in from constraints) for document d. Marginalizing the probit regression auxiliary variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ is simple in the is-a hierarchy case because they can simply be ignored. The remaining latent variables (those that are not collapsed out) are the topic indicators $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\dots,D}$, the probit regression parameters $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}$, the auxiliary variables $\mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\dots,D}$, the global topic proportions $\boldsymbol{\beta}$, and the concentration parameters α, α', and γ.

15.3.1 Gibbs Sampler

Let $\mathbf{a}$ be the set of all probit regression auxiliary variables, $\mathbf{w}$ the set of all words, $\boldsymbol{\eta}$ the set of all regression coefficients, and $\mathbf{z} \backslash z_{n,d}$ the set $\mathbf{z}$ with element $z_{n,d}$ removed.

First we consider the conditional distribution of $z_{n,d}$ (the assignment variable for each word $n = 1, \dots, N_d$ in documents $d = 1, \dots, D$). Following the factorization of the model (refer again to Figure 15.1), we can write

$$\begin{aligned} & p\left(z_{n,d} \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \\ & \quad \propto \prod_{l \in \mathcal{L}_d} p\left(a_{l,d} \mid \mathbf{z}, \boldsymbol{\eta}_l\right) p\left(z_{n,d} \mid \mathbf{z}_d \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \alpha, \boldsymbol{\beta}, \gamma\right). \end{aligned}$$

The product is only over the subset of labels $\mathcal{L}_d$ which have been observed for document d. By isolating terms that depend on $z_{n,d}$ and absorbing all other terms into a normalizing constant as in Griffiths and Steyvers (2004) we find

$$\begin{aligned} & p\left(z_{n,d} = k \mid \mathbf{z} \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \\ & \quad \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod\nolimits_{l \in \mathcal{L}_d} \exp\left\{ - \frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2} \right\}, \end{aligned} \tag{15.3}$$

where $c^{k,-(n,d)}_{v,d}$ is the number of words of type v in document d assigned to topic k omitting the

nth word of document d. The subscript $(\cdot)$ indicates to sum over the range of the replaced variable, i.e., $c^{k,-(n,d)}_{w_{n,d},(\cdot)} = \sum_d c^{k,-(n,d)}_{w_{n,d},d}$. Here, $\mathcal{L}_d$ is the set of labels which are observed for document d. We sample from (15.3) by first enumerating $z_{n,d}$ and then normalizing.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}), \tag{15.4}$$

$\boldsymbol{\eta}_l$ for $l \in \mathcal{L}$. Given that $\boldsymbol{\eta}_l$ and $a_{l,d}$ are distributed normally, the posterior distribution of $\boldsymbol{\eta}_l$ is normally distributed with mean $\hat{\boldsymbol{\mu}}_l$ and covariance $\hat{\boldsymbol{\Sigma}}$,where

$$\hat{\boldsymbol{\mu}}_l = \hat{\boldsymbol{\Sigma}} \left(\mathbf{1}\frac{\mu}{\sigma} + \bar{\mathbf{Z}}^T \mathbf{a}_l \right) \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}\sigma^{-1} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}.$$

Here $\bar{\mathbf{Z}}$ is a $D \times K$ matrix such that row d of $\bar{\mathbf{Z}}$ is $\bar{\mathbf{z}}_d$, and $\mathbf{a}_l = [a_{l,1}, a_{l,2}, \ldots, a_{l,D}]^T$. The simplicity of this conditional distribution follows from the choice of probit regression (Albert and Chib, 1993); the specific form of the update is a standard result from Bayesian normal linear regression (Gelman et al., 2004). It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution (Albert and Chib, 1993) (see also the Appendix).

$$\begin{aligned} &p\left(a_{l,d} \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\eta}\right) \\ &\quad \propto \begin{cases} \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right) \mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \\ \exp\left\{-\frac{1}{2}\left(a_{l,d} - \boldsymbol{\eta}_l^T \bar{\mathbf{z}}_d\right)\right\} \mathbb{I}\left(a_{l,d} y_{l,d} > 0\right), & y_{\mathrm{pa}(l),d} = 1. \end{cases} \end{aligned}$$

HSLDA employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than fixed a priori). This has been shown to improve the quality and stability of inferred topics (Wallach et al., 2009). Sampling $\boldsymbol{\beta}$, the vector of global topic proportions, can be done using the "direct assignment" method of Teh et al. (2006):

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'\right). \tag{15.5}$$

Here, $m_{d,k}$ are additional auxiliary variables that are introduced by the direct assignment method to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m, \tag{15.6}$$

where $s\left(n, m\right)$ denotes Stirling numbers of the first kind. The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

It remains now to show that HSLDA works. To do so we demonstrate results from modeling real-world datasets in the clinical and web retail domains. These results provide evidence that the two views (text and labels) mutually benefit multi-label classification. That is, modeling the joint is better than learning topic models and hierarchical classifiers independently.

15.4 Example Applications

15.4.1 Hospital Discharge Summaries and ICD-9 Codes

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. In the U.S., diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes.

A specific example of this involves labeling of hospital discharge summaries. These summaries are authored by clinicians to summarize patient hospitalization courses. They typically contain a record of patient complaints, findings, and diagnoses, along with treatment and hospital course. The kind of text one might expect to find in such a discharge summary is illustrated by this made-up snippet:

> History of Present Illness: Mrs. Carmen Sandiego is a 62-year-old female with a past medical history significant for diabetes, hypertension, hyperlipidemia, afib, status post MI in 5/2010 and cholecystectomy in 3/2009. The patient presented to the ED on 7/11/2011 with a right sided partial facial hemiparesis along with mild left arm weakness. The patient was admitted to the Neurology service and underwent a workup for stroke given her history of MI and many cardiovascular risk factors ...

For each hospitalization, trained medical coders review the information in the discharge summary and assign a series of diagnoses codes. Coding follows the ICD-9-CM controlled terminology, an international diagnostic classification for epidemiological, health management, and clinical purposes.[1] These ICD-9 codes are organized in a rooted-tree structure with each edge representing an is-a relationship between parent and child such that the parent diagnosis subsumes the child diagnosis. For example, the code for "Pneumonia due to adenovirus" is a child of the code for "Viral pneumonia," where the former is a type of the latter. A representative sub-tree of the ICD-9 code tree is shown in Figure 15.2. It is worth noting that the coding can be noisy. Human coders sometimes disagree (Cha, 2007), tend to be more specific than sensitive in their assignments (Birman-Deych et al., 2005), and sometimes make mistakes (Farzandipour et al., 2010).

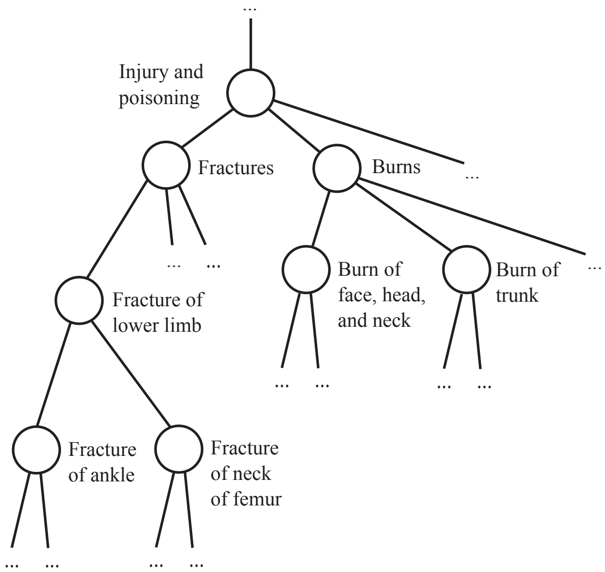


FIGURE 15.2
An illustration of a portion of the ICD9 hierarchy.

An automated process would ideally produce more complete and accurate diagnosis lists. The

[1] See http://www.cdc.gov/nchs/icd/icd9cm.htm.

task of automatic ICD-9 coding has been investigated in the clinical domain. Methods used to solve this problem (besides HSLDA) range from applying manually derived coding rules to applications of online rule learning approaches (Crammer et al., 2007; Goldstein et al., 2007; Farkas and Szarvas, 2008). Many classification schemes have been applied to this problem: K-nearest neighbor, naive Bayes, support vector machines, Bayesian Ridge Regression, as well as simple keyword mappings, all with promising results (Larkey and Croft, 1995; Ribeiro-Neto et al., 2001; Pakhomov et al., 2006; Lita et al., 2008).

The specific dataset we report results for in this chapter was gathered from the New York-Presbyterian Hospital clinical data warehouse. It consists of 6000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing a portion of the discharges from the hospital in 2009. All included discharge summaries had associated ICD-9 Codes. Summaries have 8.39 associated ICD-9 codes on average (std dev=5.01) and contain an average of 536.57 terms after preprocessing (std dev=300.29). We split our dataset into 5000 discharge summaries for training and 1000 for testing.

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with the highest document frequency (exclusive of names, places, and other identifying numbers). The study was approved by the Institutional Review Board and follows HIPAA (Health Insurance Portability and Accountability Act) privacy guidelines.

Here HSLDA is evaluated as a way to understand and model the relationship between a discharge summary and the ICD-9 codes that should be assigned to it. We show promising results for automatically assigning ICD-9 codes to hospital discharge records.

15.4.2 Product Descriptions and Catalogs

Many web-retailers store and organize their catalog of products in a mulitply-rooted hierarchy in addition to providing textual product descriptions for most products. Products can be discovered by users through free-text search and product category exploration. Top-level product categories are displayed on the front page of the website and lower-level categories can be discovered by choosing one of the top-level categories. Products can exist in multiple locations in the hierarchy.

Amazon.com is one such retailer. Its product categorization data is available as part of the Stanford Network Analysis Platform (SNAP) dataset (SNA, 2004). A representative sub-tree of the Amazon.com DVD product category tree is shown in Figure 15.3. Product descriptions were obtained separately from the Amazon.com website directly. Once such description is

> Winner of five Academy Awards, including Best Picture and Best Director, *The Deer Hunter* is simultaneously an audacious directorial conceit and one of the greatest films ever made about friendship and the personal impact of war. Like *Apocalypse Now*, it's hardly a conventional battle film—the soldier's experience was handled with greater authenticity in *Platoon*—but its depiction of war on an intimate scale packs a devastatingly dramatic punch ...

We study the collection of DVDs in the product catalog specifically. The resulting dataset contains 15,130 product descriptions for training and 1000 for testing. The product descriptions consist of 91.89 terms on average (std dev=53.08). Overall, there are 2,691 unique categories. Products are assigned on average 9.01 categories (std dev=4.91). The vocabulary consists of the most frequent 30,000 words omitting stopwords.

HSLDA is used here to understand and model the relationship between the product text description and the products' positioning in the product hierarchy. We show how to automatically situate a product in a hierarchical product catalog.

[2] See http://www.nltk.org.

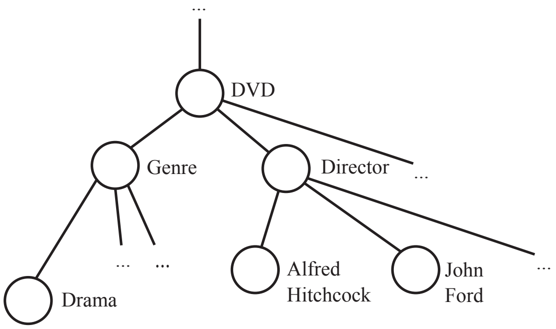


FIGURE 15.3
An illustration of a portion of the Amazon product hierarchy.

15.4.3 Comparison Models

We compare HSLDA to two closely related models. The comparison models are SLDA with independent regressors (hierarchical constraints on labels ignored, i.e., the regression is not conditional) and HSLDA fit by first performing LDA, then fitting probit regressors that respect the conditional label hierarchy (rather than jointly inferring the topics and the regression coefficients). These models were chosen because they are the strongest available competitors and because they highlight several pedagogical aspects of HSLDA, including performance in the absence of hierarchical constraints, the effect of the combined inference, and regression performance attributable solely to the hierarchical constraints.

SLDA with independent regressors is the most salient comparison model for our work. The distinguishing factor between HSLDA and SLDA is the additional structure imposed on the label space, a distinction that in developing HDSLA we hypothesized would result in a difference in predictive performance.

The second comparison model, HSLDA fit by performing LDA first followed by performing inference over the hierarchically constrained label space, does not allow the responses to influence the topics inferred by LDA. Combined inference has been shown to improve performance in SLDA (Blei and McAuliffe, 2008). This comparison model does not examine the value of utilizing the structured nature of the label space; instead, it highlights the benefit of combined inference over both the documents and the label space.

For all three models, particular attention was given to the settings of the prior parameters (μ, σ) for the regression coefficients (ν_l). These parameters implement an important form of regularization in HSLDA. In the setting where there are no negative labels, a Gaussian prior over the regression parameters with a negative mean implements a prior belief that missing labels are likely to be negative. Thus, we show model performance for all three models with a range of values for μ, the mean prior parameter for regression coefficients ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

The number of topics for all models was set to $K = 50$, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were all chosen to be gamma with a shape parameter of 1 and a scale parameter of 1000. Different values of K corresponding to different numbers of topics were explored, however, the results that we show in the following are not substantially changed in character. As is usual in mixed membership models, there is an ideal number of topics that should be used for out-of-sample prediction tasks, however, a full model-selection search varying topic cardinality was not performed for these datasets.

15.4.4 Evaluation and Results

We are particularly interested in the predictive performance on held-out data. Prediction performance was measured with standard metrics—sensitivity (true positive rate) and 1-specificity (false positive rate).

In each case the gold standard for testing was derived from the test data. To make the comparison as antagonistic to HSLDA as possible (relative to the other models), in evaluation only, ancestors of observed nodes in the label hierarchy were ignored, observed nodes were considered positive, and descendants of observed nodes were assumed to be negative. Note that this is different from our treatment of the observations during inference where we marginalize over possible settings of unobserved labels. For instance, as the SLDA model does not enforce hierarchical label constraints, when we consider only observed nodes we penalize HSLDA. This is because the is-a hierarchical constraints say that the ancestors of positively labeled nodes must also be positive, which the SLDA model cannot guarantee. Another antagonism of this gold standard is that it is likely to inflate the number of false positives because the labels applied to any particular document are usually not as complete as they could be. ICD-9 codes, for instance, are known to lack sensitivity and their use as a gold standard could lead to correctly positive predictions being labeled as false positives (Birman-Deych et al., 2005). However, given that the label space is often large (as in our examples), it is a reasonable assumption that erroneous false positives should not skew results significantly.

Predictive performance in HSLDA is evaluated by computing

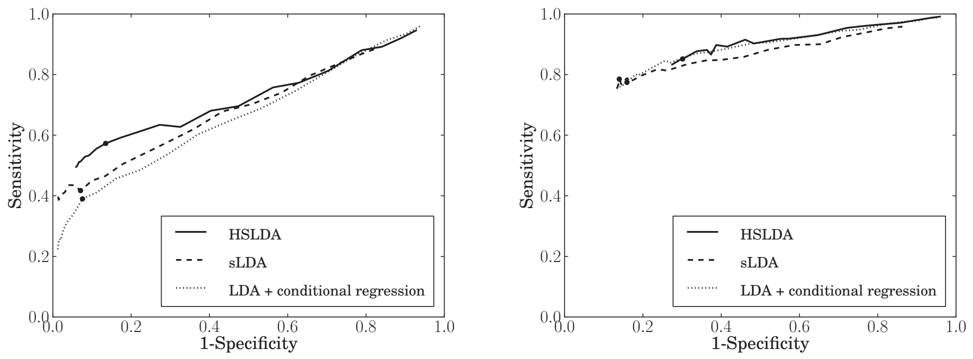
$$p\left(y_{l,\hat{d}} \mid w_{1:N_{\hat{d}},\hat{d}}, w_{1:N_d,1:D}, y_{l\in\mathcal{L},1:D}\right)$$

for each test document $\hat{d}$ for each observed label $y_{l,\hat{d}}$ (given the test document words). For efficiency, the expectation of this probability distribution was approximated in the following way: Expectations of $\bar{\mathbf{z}}_{\hat{d}}$ and $\boldsymbol{\eta}_l$ were estimated with samples from the posterior. Fixing these expectations, we performed Gibbs sampling over the hierarchy to acquire predictive samples for the documents in the test set. The true positive rate was calculated as the average expected labeling for gold standard positive labels. The false positive rate was calculated as the average expected labeling for gold standard negative labels.

As sensitivity and specificity can always be traded off, we examined sensitivity for a range of values for two different parameters—the prior means for the regression coefficients and the threshold for the auxiliary variables. The goal in this analysis was to evaluate the performance of these models subject to more or less stringent requirements for predicting positive labels. These two parameters have important related functions in the model. The prior mean in combination with the auxiliary variable threshold together encode the strength of the prior belief that unobserved labels are likely to be negative. Effectively, the prior mean applies negative pressure to the predictions and the auxiliary variable threshold determines the cutoff. For each model type, separate models were fit for each value of the prior mean of the regression coefficients. This is a proper Bayesian sensitivity analysis. In contrast, to evaluate predictive performance as a function of the auxiliary variable threshold, a single model was fit for each model type and prediction was evaluated based on predictive samples drawn subject to different auxiliary variable thresholds. These methods are significantly different since the prior mean is varied prior to inference, and the auxiliary variable threshold is varied following inference.

Figure 15.4(a) demonstrates the performance of the model on the clinical data as a ROC curve varying μ. For instance, a hyperparameter setting of $\mu = -1.6$ yields the following performance: the full HSLDA model had a true positive rate of 0.57 and a false positive rate of 0.13, the SLDA model had a true positive rate of 0.42 and a false positive rate of 0.07, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.39 and a false positive rate of 0.08. These points are highlighted in Figure 15.4(a). Note that the figure is somewhat misleading because for any one value of μ, HSLDA outperforms the comparison models by a relatively large margin.

These results indicate that the full HSLDA model predicts more of the correct labels at a cost of an increase in the number of false positives relative to the comparison models. However, as shown in Figure 15.4(a), HSLDA outperforms no worse than the comparison models across the full range of specificities.



(a) Clinical data performance

(b) Retail product performance

FIGURE 15.4
ROC curves for HSLDA out-of-sample label prediction varying μ, the prior mean of the regression parameters. In both figures, solid is HSLDA, dashed are independent regressors + SLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

Example topics (as word lists) learned for the discharge data are given below. These word lists are computed by sorting terms in decreasing order based on their probability under a given topic.

Topic 1	**Topic 2**
MASS	WOUND
CANCER	FOOT
RIGHT	CELLULITIS
BREAST	ULCER
CHEMOTHERAPY	LEFT
METASTATIC	ERYTHEMA
LEFT	PAIN
LYMPH	SWELLING
TUMOR	SKIN
BIOPSY	RIGHT
CARCINOMA	ABSCESS
LUNG	LEG
CHEMO	OSTEOMYELITIS
ADENOCARCINOMA	TOE
NODE	DRAINAGE

These topics closely correspond to common clinical concepts, namely cancers of the thorax and wounds common to diabetics suffering from poor peripheral circulation. Evaluations of the subject coherence of these topics relative to baselines are ongoing, but early results suggest positive findings similar to those reported for other supervised LDA models.

Figure 15.4(b) demonstrates the performance of the model on the retail product data as an ROC curve also varying μ. For instance, a hyperparameter setting of $\mu = -2.2$ yields the following performance: the full HSLDA model had a true positive rate of 0.85 and a false positive rate of 0.30, the SLDA model had a true positive rate of 0.78 and a false positive rate of 0.14, and the HSLDA model where LDA and the regressions were fit separately had a true positive rate of 0.77 and a false positive rate of 0.16. These results follow a similar pattern to the clinical data. These points are highlighted in Figure 15.4(b).

Example topics (as word lists) learned for the Amazon.com data are given below. These word lists were also computed by sorting terms in decreasing order based on their probability under a given topic.

Topic 1	**Topic 2**
SERIES	BASEBALL
EPISODES	TEAM
SHOW	GAME
SEASON	PLAYERS
EPISODE	BASKETBALL
FIRST	SPORT
TELEVISION	SPORTS
SET	NEW
TIME	PLAYER
TWO	SEASON
SECOND	LEAGUE
ONE	FOOTBALL
CHARACTERS	STARS
DISC	FANS
GUEST	FIELD

Figure 15.5 shows the predictive performance of HSLDA relative to the two comparison models on the clinical dataset as a function of the auxiliary variable threshold. For low values of the auxiliary variable threshold, the models predict labels in a more sensitive and less specific manner, creating the points in the upper right corner of the ROC curve. As the auxiliary variable threshold is increased, the models predict in a less sensitive and more specific manner, creating the points in the lower left hand corner of the ROC curve. HSLDA with full joint inference outperforms SLDA with independent regressors as well as HSLDA with separately trained regression.

15.5 Related Work

HSLDA does not, of course, stand alone. Models for structured labeling of bag-of-words data can be designed in a number of different ways.

As shown in Section 15.4, SLDA can be used to solve this kind of problem directly, however, doing so requires ignoring the hierarchical dependencies amongst the labels. Other models that incorporate LDA and supervision that could also be used to solve this problem include LabeledLDA (Ramage et al., 2009) and DiscLDA (Lacoste-Julien et al.). Various applications of these models to computer vision and document networks have been explored (Wang et al., 2009; Chang

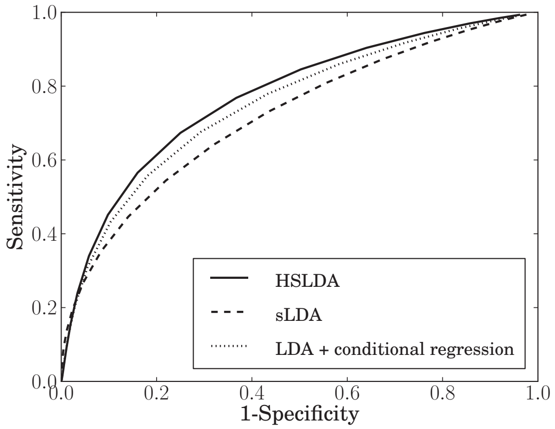


FIGURE 15.5
ROC curve for out-of-sample ICD-9 code prediction varying auxiliary variable threshold. $\mu = -1.0$ for all three models in this figure.

and Blei, 2010). None of these models, however, leverage dependency structure in the label space.

In other non-LDA-based related work, researchers have classified documents into hierarchies (a closely related task) using naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets and small label spaces, and has focused on single label classification without a model of documents such as LDA (McCallum et al., 1999; Dumais and Chen, 2000; Koller and Sahami, 1997; Chakrabarti et al., 1998).

15.6 Discussion

The SLDA model family, of which HSLDA is a member, can be understood in two different ways. One way is to see it as a family of topic models that improve on the topic modeling performance of LDA via the inclusion of observed supervision. An alternative, complementary way is to see it as a set of models that can predict labels for bag-of-word data. A large diversity of problems can be expressed as label prediction problems for bag-of-word data. A surprisingly large amount of data possess structured labels, either hierarchically constrained or otherwise. HSLDA directly addresses this kind of data and works well in practice. That it outperforms more straightforward approaches should be of interest to practitioners.

There are many kinds of problems that have the same characteristics as this: any data that consists of free text that has been partially or completely categorized by human editors; more specifically, any bag-of-words data that has been, at least in part, categorized. Examples include, but are not limited to, webpages and curated hierarchical directories of the same (DMO, 2002), product descriptions and catalogs, (e.g., AMA (2011) as available from SNA (2004)) and patient records and diagnosis codes assigned to them for bookkeeping and insurance purposes. The model we cover in this chapter shows one way to combine these two sources of information into a single model allowing one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more.

Extensions to this work include a nonparametric Bayesian extension with unbounded topic cardinality and relaxations to different kinds of label structure. Unbounded topic cardinality variants pose interesting inference challenges. Imposing different kinds of label structure constraints is possible within this framework but requires relaxing some of the assumptions we made in deriving the sampling distributions for HSLDA inference.

Appendix

Probit Regression

For reasons that are somewhat obscure, statisticians tend to use probit regression for binary classification whereas machine learners tend to use logistic regression. The "probit" function is the inverse of the normal cumulative distribution function (cdf). We denote the normal cdf function $\Phi(x; \mu, \sigma^2)$ with μ the mean, σ^2 the variance, and x the argument.

The range of the normal cdf is $(0, 1)$, which means that it can be interpreted as a probability. For instance, one can construct a generalized linear classification model (a "probit regression model") of the form

$$P(y_i = 1) = \Phi(x_i^T \beta; 0, \sigma^2). \tag{15.7}$$

Depending on convention (i.e., binary y_i represented as $\{1, 0\}$ or $\{1, -1\}$), the probability of y_i being labeled the opposite way is $P(y_i = -1)$ or $P(y_i = 0) = 1 - P(y_i = 1)$. Here x_i is a vector of covariates, β is a vector of weights, and y_i is a single, binary valued response. The close relationship between regression and classification is in full display here: probit regression is a "generalized linear *regression* model" as well as a "binary classifier."

In this model we would like to use labeled training data, $\{x_i, y_i\}_{i=1}^N$ to "learn" the value of β and then to use this value to predict the value of $y_{N+1}|x_{N+1}, \beta$. Being Bayesian about inference means that we will average over the posterior distribution of β when making predictions. This means that we want to draw samples from the posterior distribution of $\beta|\{x_i, y_i\}_{i=1}^N$. To do this efficiently one can introduce a set of auxiliary variables $\{u_i\}_{i=1}^N$.

By auxiliary variables we mean that such variables will be used as an intermediary for purposes of efficiency but will otherwise be uninteresting. They are variables introduced into a model in order to make inference easier but whose existence does not change the distribution of interest. Auxiliary variables for slice sampling are one particularly clever use of auxiliary variables. The auxiliary variable trick in probit regression is another.

For the purposes of exposition, forget about the i index and focus on a single instance y, x, and u. The argument we make will hold for all by simply reintroducing subscripts.

To start, let's propose a factorized joint distribution for these quantities

$$P(y, x, u) = P(y|u)P(u|x, \beta). \tag{15.8}$$

Straight away, one can see why this auxiliary variable scheme works. By the law of total probability we have

$$P(y, x) = \int P(y, x, u)du = \int P(y|u)P(u|x, \beta)du. \tag{15.9}$$

So, if by some means we generate S samples $\{u^{(s)}, y^{(s)}, x^{(s)}\}_{s=1}^S \sim P(y, x, u)$, we know that marginalizing u out (i.e., disregarding its value) we get samples $\{y^{(s)}, x^{(s)}\}_{s=1}^S \sim P(y, x)$.

We haven't specified the most important part of the auxiliary variable sampling scheme yet, namely, what $P(y|u)$ and $P(u|x, \beta)$ are. Let us try $y = \text{sign}(u)$ and $u \sim N(x^T\beta, \sigma^2)$. These choices are nice in a particular way. First let us verify that the marginalization of u out of this model results in the model specification in Equation (15.7):

$$\begin{aligned} P(y=1|x,\beta) &= \int P(y=1|u)P(u|x,\beta)du \\ &= \int \mathbb{I}(u>0)N(u;x^T\beta,\sigma^2)du \\ &= \int_0^\infty N(u;x^T\beta,\sigma^2)du \\ &= 1-\Phi(0;x^T\beta,\sigma^2) \\ &= \Phi(x^T\beta;0,\sigma^2), \end{aligned}$$

where the last line comes from the fact that for symmetric distributions like the normal distribution, $\Phi(x^T\beta;0,\sigma^2)=1-\Phi(-x^T\beta;0,\sigma^2)$, and the mean of a normal cdf can be translated arbitrarily, i.e., $\Phi(-x^T\beta;0,\sigma^2)=\Phi(0;x^T\beta,\sigma^2)$ (which comes from adding the offset $x^T\beta$ to the cdf argument and mean).

Having established the fact that for a particular sort of auxiliary variable choice, we get the same probit model as we wanted, why is this choice nice?

Well, it comes down to sampling β, u, and y. Generally, sampling β in the model without auxiliary variables will require hybrid Monte Carlo (HMC) or Metropolis-Hastings of some sort. Gibbs sampling often comes with substantial benefits. By making this choice of auxiliary variable, the conditional distribution of u_i given everything else is proportional to a truncated normal distribution, a distribution that is, by nature of its commonness, relatively straightforward to sample from. The big benefit, though, acrues from the posterior form for sampling β. With the u's "observed" (as they would be in a Gibbs sampler), the posterior distribution of β (for typical choices of prior) is precisely the same as that for linear regression, perhaps the most well-studied model in statistics. In that case, sampling β from its posterior distribution is quite simple usually, and certainly more so than sampling β without the u auxiliary variables.

The extension to the multivariate HSLDA setting is straightforward and follows this line of reasoning precisely. An extended discussion of the techniques suggested here and the multivariate generalization can be found in Gelman et al. (2004).

References

(2002). DMOZ open directory project. http://www.dmoz.org/.

(2004). Stanford network analysis platform. http://snap.stanford.edu/.

(2007). The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous.

(2011). Amazon, Inc. http://www.amazon.com/.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669.

Birman-Deych, E., Waterman, A. D., Yan, Y., Nilasena, D. S., Radford, M. J., and Gage, B. F. (2005). Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care* 43: 480–5.

Blei, D. M. and McAuliffe, J. (2008). Supervised topic models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. (eds), *Advances in Neural Information Processing Systems 20*. Cambridge, MA: The MIT Press, 121–128.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Chakrabarti, S., Dom, B., Agrawal, R., and Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal* 7: 163–178.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks. *Annals of Applied Statistics* 4: 124–150.

Crammer, K., Dredze, M., Ganchev, K., Talukdar, P., and Carroll, S. (2007). Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 129–136.

Dumais, S. and Chen, H. (2000). Hierarchical classification of web content. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*. New York, NY, USA: ACM, 256–263.

Farkas, R. and Szarvas, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics* 9: S10.

Farzandipour, M., Sheikhtaheri, A., and Sadoughi, F. (2010). Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision. *International Journal of Information Management* 30: 78–84.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition.

Goldstein, I., Arzumtsyan, A., and Uzuner, Ö. (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *AMIA Annual Symposium Proceedings* 2007: 279–283.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science (PNAS)* 101: 5228–5235.

Koller, D. and Sahami, M. (1997). Hierarchically Classifying Documents Using Very Few Words. Tech. report 1997-75, Stanford InfoLab, previous number = SIDL-WP-1997-0059.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds), *Advances in Neural Information Processing Systems 21*. Red Hook, NY: Curran Associates, Inc., 897–904.

Larkey, L. and Croft, B. (1995). Automatic Assignment of ICD9 Codes to Discharge Summaries. Tech. report, University of Massachussets.

Lita, L. V., Yu, S., Niculescu, S., and Bi, J. (2008). Large scale diagnostic code classification for medical patient records. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*. Asian Federation for Natural Language Processing, 877–882.

McCallum, A., Nigam, K., Rennie, J. D. M., and Seymore, K. (1999). Building domain-specific search engines with machine learning techniques. In *Proceedings of the 16th International Joint Conference on Artifical Intelligence - Volume 2 (IJCAI '99)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 662–667.

Pakhomov, S., Buntrock, J., and Chute, C. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association (JAMIA)* 13: 516–525.

Perotte, A., Barlett, N., Elhadad, N., and Wood, F. (2011). Hierarchically supervised latent Dirichlet allocation. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P., Pereira, F. C. N., and Weinberger, K. Q. (eds), *Advances in Neural Information Processings Systems 24*. Red Hook, NY: Curran Associates, Inc., 2609–2617.

Ramage, D., Hall, D., Nallapati, R. M., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 248–256.

Ribeiro-Neto, B., Laender, A., and Lima, L. D. (2001). An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology* 52: 391–401.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101: 1566–1581.

Wallach, H. M., Mimno, D., and McCallum, A. (2009). Rethinking LDA: Why priors matter. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds), *Advances in Neural Information Processing Systems 22*. Red Hook, NY: Curran Associates, Inc., 1973–1981.

Wang, C., Blei, D. M., and Fei-Fei, L. (2009). Simultaneous image classification and annotation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*. Los Alamitos, CA, USA: IEEE Computer Society, 1903–1910.