Author Name

Book title goes here

Foreward

I am delighted to introduce the first book on Multimedia Data Mining. When I came to know about this book project undertaken by two of the most active young researchers in the field, I was pleased that this book is coming in early stage of a field that will need it more than most fields do. In most emerging research fields, a book can play a significant role in bringing some maturity to the field. Research fields advance through research papers. In research papers, however, only a limited perspective could be provided about the field, its application potential, and the techniques required and already developed in the field. A book gives such a chance. I liked the idea that there will be a book that will try to unify the field by bringing in disparate topics already available in several papers that are not easy to find and understand. I was supportive of this book project even before I had seen any material on it. The project was a brilliant and a bold idea by two active researchers. Now that I have it on my screen, it appears to be even a better idea.

Multimedia started gaining recognition in 1990s as a field. Processing, storage, communication, and capture and display technologies had advanced enough that researchers and technologists started building approaches to combine information in multiple types of signals such as audio, images, video, and text. Multimedia computing and communication techniques recognize correlated information in multiple sources as well as insufficiency of information in any individual source. By properly selecting sources to provide complementary information, such systems aspire, much like human perception system, to create a holistic picture of a situation using only partial information from separate sources.

Data mining is a direct outgrowth of progress in data storage and processing speeds. When it became possible to store large volume of data and run different statistical computations to explore all possible and even unlikely correlations among data, the field of data mining was born. Data mining allowed people to hypothesize relationships among data entities and explore support for those. This field has been put to applications in many diverse domains and keeps getting more applications. In fact many new fields are direct outgrowth of data mining and it is likely to become a powerful computational tool.

Preface

Approximately 17 million people in the USA (6% of the population) and 140 million people worldwide (this number is expected to rise to almost 300 million by the year 2025) suffer from *diabetes mellitus*. Currently, there a few dozens of commercialised devices for detecting blood glucose levels [1]. However, most of them are invasive. The development of a noninvasive method would considerably improve the quality of life for diabetic patients, facilitate their compliance for glucose monitoring, and reduce complications and mortality associated with this disease. Noninvasive and continuous monitoring of glucose concentration in blood and tissues is one of the most challenging and exciting applications of optics in medicine. The major difficulty in development and clinical application of optical noninvasive blood glucose sensors is associated with very low signal produced by glucose molecules. This results in low sensitivity and specificity of glucose monitoring by optical methods and needs a lot of efforts to overcome this difficulty.

A wide range of optical technologies have been designed in attempts to develop robust noninvasive methods for glucose sensing. The methods include infrared absorption, near-infrared scattering, Raman, fluorescent, and thermal gradient spectroscopies, as well as polarimetric, polarization heterodyning, photonic crystal, optoacoustic, optothermal, and optical coherence tomography (OCT) techniques [1-31].

For example, the polarimetric quantification of glucose is based on the phenomenon of optical rotatory dispersion, whereby a chiral molecule in an aqueous solution rotates the plane of linearly polarized light passing through the solution. The angle of rotation depends linearly on the concentration of the chiral species, the pathlength through the sample, and the molecule specific rotation. However, polarization sensitive optical technique makes it difficult to measure *in vivo* glucose concentration in blood through the skin because of the strong light scattering which causes light depolarization. For this reason, the anterior chamber of the eye has been suggested as a sight well suited for polarimetric measurements, since scattering in the eye is generally very low compared to that in other tissues, and a high correlation exists between the glucose in the blood and in the aqueous humor. The high accuracy of anterior eye chamber measurements is also due to the low concentration of optically active aqueous proteins within the aqueous humor.

On the other hand, the concept of noninvasive blood glucose sensing using the scattering properties of blood and tissues as an alternative to spectral absorption and polarization methods for monitoring of physiological glucose concentrations in diabetic patients has been under intensive discussion for the last decade. Many of the considered effects, such as changing of the size, refractive index, packing, and aggregation of RBC under glucose variation, are important for glucose monitoring in diabetic patients. Indeed, at physiological concentrations of glucose, ranging from 40 to 400 mg/dl, the role of some of the effects may be modified, and some other effects, such as glucose penetration inside the RBC and the followed hemoglobin glycation, may be important [30-32].

Noninvasive determination of glucose was attempted using light scattering of skin tissue components measured by a spatially-resolved diffuse reflectance or NIR frequency-domain reflectance techniques. Both approaches are based on change in glucose concentration, which affects the refractive index mismatch between the interstitial fluid and tissue fibers, and hence reduces scattering coefficient. A glucose clamp experiment showed that reduced scattering coefficient measured in the visible range qualitatively tracked changes in blood glucose concentration for the volunteer with diabetes studied.

Contributors

Michael Aftosmis
NASA Ames Research Center
Moffett Field, California

Pratul K. Agarwal
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Sadaf R. Alam
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Gabrielle Allen
Louisiana State University
Baton Rouge, Louisiana

Martin Sandve Alnæs
Simula Research Laboratory and
University of Oslo, Norway
Norway

Steven F. Ashby
Lawrence Livermore National
Laboratory
Livermore, California

David A. Bader
Georgia Institute of Technology
Atlanta, Georgia

Benjamin Bergen
Los Alamos National Laboratory
Los Alamos, New Mexico

Jonathan W. Berry
Sandia National Laboratories
Albuquerque, New Mexico

Martin Berzins
University of Utah
Salt Lake City, Utah

Abhinav Bhatele
University of Illinois
Urbana-Champaign, Illinois

Christian Bischof
RWTH Aachen University
Germany

Rupak Biswas
NASA Ames Research Center
Moffett Field, California

Eric Bohm
University of Illinois
Urbana-Champaign, Illinois

James Bordner
University of California, San Diego
San Diego, California

George Bosilca
University of Tennessee
Knoxville, Tennessee

Greg L. Bryan
Columbia University
New York, New York

Marian Bubak
AGH University of Science and
Technology

Kraków, Poland

Andrew Canning
Lawrence Berkeley National Laboratory
Berkeley, California

Jonathan Carter
Lawrence Berkeley National Laboratory
Berkeley, California

Zizhong Chen
Jacksonville State University
Jacksonville, Alabama

Joseph R. Crobak
Rutgers, The State University of New Jersey
Piscataway, New Jersey

Roxana E. Diaconescu
Yahoo! Inc.
Burbank, California

Peter Diener
Louisiana State University
Baton Rouge, Louisiana

Jack J. Dongarra
University of Tennessee, Knoxville, Oak Ridge National Laboratory, and
University of Manchester

John B. Drake
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Kelvin K. Droegemeier
University of Oklahoma
Norman, Oklahoma

Stéphane Ethier
Princeton University
Princeton, New Jersey

Christoph Freundl
Friedrich–Alexander–Universität
Erlangen, Germany

Karl Fürlinger
University of Tennessee
Knoxville, Tennessee

Al Geist
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Michael Gerndt
Technische Universität München
Munich, Germany

Tom Goodale
Louisiana State University
Baton Rouge, Louisiana

Tobias Gradl
Friedrich–Alexander–Universität
Erlangen, Germany

William D. Gropp
Argonne National Laboratory
Argonne, Illinois

Robert Harkness
University of California, San Diego
San Diego, California

Albert Hartono
Ohio State University
Columbus, Ohio

Thomas C. Henderson
University of Utah
Salt Lake City, Utah

Bruce A. Hendrickson
Sandia National Laboratories
Albuquerque, New Mexico

Alfons G. Hoekstra
University of Amsterdam
Amsterdam, The Netherlands

Philip W. Jones
Los Alamos National Laboratory
Los Alamos, New Mexico

Laxmikant Kalé
University of Illinois
Urbana-Champaign, Illinois

Shoaib Kamil
Lawrence Berkeley National Laboratory
Berkeley, California

Cetin Kiris
NASA Ames Research Center
Moffett Field, California

Uwe Küster
University of Stuttgart
Stuttgart, Germany

Julien Langou
University of Colorado
Denver, Colorado

Hans Petter Langtangen
Simula Research Laboratory and
University of Oslo, Norway

Michael Lijewski
Lawrence Berkeley National Laboratory
Berkeley, California

Anders Logg
Simula Research Laboratory and
University of Oslo, Norway

Justin Luitjens
University of Utah
Salt Lake City, Utah

Kamesh Madduri
Georgia Institute of Technology
Atlanta, Georgia

Kent-Andre Mardal
Simula Research Laboratory and
University of Oslo, Norway

Satoshi Matsuoka
Tokyo Institute of Technology
Tokyo, Japan

John M. May
Lawrence Livermore National Laboratory
Livermore, California

Celso L. Mendes
University of Illinois
Urbana-Champaign, Illinois

Dieter an Mey
RWTH Aachen University
Germany

Tetsu Narumi
Keio University
Japan

Michael L. Norman
University of California, San Diego
San Diego, California

Boyana Norris
Argonne National Laboratory
Argonne, Illinois

Yousuke Ohno
Institute of Physical and Chemical Research (RIKEN)
Kanagawa, Japan

Leonid Oliker
Lawrence Berkeley National Laboratory
Berkeley, California

Brian O'Shea
Los Alamos National Laboratory
Los Alamos, New Mexico

Christian D. Ott
University of Arizona
Tucson, Arizona

James C. Phillips
University of Illinois
Urbana-Champaign, Illinois

Simon Portegies Zwart
University of Amsterdam,
Amsterdam, The Netherlands

Thomas Radke
Albert-Einstein-Institut
Golm, Germany

Michael Resch
University of Stuttgart
Stuttgart, Germany

Daniel Reynolds
University of California, San Diego
San Diego, California

Ulrich Rüde
Friedrich–Alexander–Universität
Erlangen, Germany

Samuel Sarholz
RWTH Aachen University
Germany

Erik Schnetter
Louisiana State University
Baton Rouge, Louisiana

Klaus Schulten
University of Illinois
Urbana-Champaign, Illinois

Edward Seidel
Louisiana State University
Baton Rouge, Louisiana

John Shalf
Lawrence Berkeley National Laboratory
Berkeley, California

Bo-Wen Shen
NASA Goddard Space Flight Center
Greenbelt, Maryland

Ola Skavhaug
Simula Research Laboratory and
University of Oslo, Norway

Peter M.A. Sloot
University of Amsterdam
Amsterdam, The Netherlands

Erich Strohmaier
Lawrence Berkeley National Laboratory
Berkeley, California

Makoto Taiji
Institute of Physical and Chemical Research (RIKEN)
Kanagawa, Japan

Christian Terboven
RWTH Aachen University,
Germany

Mariana Vertenstein
National Center for Atmospheric Research
Boulder, Colorado

Rick Wagner
University of California, San Diego
San Diego, California

Daniel Weber
University of Oklahoma
Norman, Oklahoma

James B. White, III
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Terry Wilmarth
University of Illinois
Urbana-Champaign, Illinois

List of Figures

1.1 Hierarchically supervised latent Dirichlet allocation (HSLDA) graphical model. 7
1.2 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. 15
1.3 ROC curve for out-of-sample ICD-9 code prediction varying auxiliary variable threshold. $\mu = -1.0$ for all three models in this figure. 16
1.4 Probit model, no auxiliary variables 19
1.5 CDF of $N(0, 1)$ 19
1.6 Probit model 20

List of Tables

Contents

I This is a Part **1**

1 Mixed Membership Classification **3**
Frank Wood and Adler Perotte
1.1 Introduction . . . 3
1.1.1 Background . . . 4
1.1.2 Intuition . . . 5
1.2 Hierarchical Supervised LDA (HSLDA) . . . 6
1.2.1 Ancestral sampling for HSLDA . . . 7
1.3 Inference . . . 9
1.3.1 Gibbs Sampler . . . 10
1.4 Example Applications . . . 11
1.4.1 Hospital Discharge Summaries and ICD-9 Codes . . . 12
1.4.2 Product Descriptions and Catalogs . . . 13
1.4.3 Comparison Models . . . 13
1.4.4 Evaluation and Results . . . 14
1.5 Discussion . . . 18
1.6 Appendix . . . 18
1.6.1 MCMC . . . 18
1.6.2 Auxiliary Variables . . . 18
1.6.3 Probit Regression . . . 18

Bibliography **23**

Index **27**

Symbol Description

α	To solve the generator maintenance scheduling, in the past, several mathematical techniques have been applied.
σ^2	These include integer programming, integer linear programming, dynamic programming, branch and bound etc.
$\sum$	Several heuristic search algorithms have also been developed. In recent years expert systems,
abc	fuzzy approaches, simulated annealing and genetic algorithms have also been tested.
$\theta\sqrt{abc}$	This paper presents a survey of the literature
ζ	over the past fifteen years in the generator
∂	maintenance scheduling. The objective is to
sdf	present a clear picture of the available recent literature
ewq	of the problem, the constraints and the other aspects of
bvcn	the generator maintenance schedule.

Part I

This is a Part

1

Mixed Membership Classification

Frank Wood

Columbia University, Department of Statistics, New York, NY 10027 USA

Adler Perotte

Columbia University, ..., New York, NY 10027

CONTENTS

1.1 Introduction 3
  1.1.1 Background 4
  1.1.2 Intuition 5
1.2 Hierarchical Supervised LDA (HSLDA) 6
  1.2.1 Ancestral sampling for HSLDA 7
1.3 Inference 9
  1.3.1 Gibbs Sampler 10
1.4 Example Applications 11
  1.4.1 Hospital Discharge Summaries and ICD-9 Codes 11
  1.4.2 Product Descriptions and Catalogs 13
  1.4.3 Comparison Models 13
  1.4.4 Evaluation and Results 14
1.5 Discussion 17
1.6 Appendix 18
  1.6.1 MCMC 18
  1.6.2 Auxiliary Variables 18
  1.6.3 Probit Regression 18

1.1 Introduction

This chapter covers classifying or, equivalently, labeling bag-of-words data. We show how to construct classifiers based on features derived from mixed membership models. We specifically develop a solution to the concrete problem of hierarchically classifying text documents. The model and techniques developed in this chapter can be used to solve any labeling or hierarchical classification problem involving data represented as bags-of-words. If, for instance, an image is represented by counts of features then it is said to be in bag-of-words representation. So, for instance, hierarchical classification of images might be tackled using the techniques developed in this chapter, although, such a particular application will not be explicitly considered.

We should agree that classification and regression are largely the same. Namely, classification is regression with discrete output. To solve classification problems we must learn from training instances, as in regression, a function $f : \mathcal{X} \to \mathcal{L}$ which maps from an input space $\mathcal{X}$ to an output (label) space $\mathcal{L}$. As an example consider the problem of identifying the authors of a set of documents for which the true author is unknown but is one of a set of possible authors. For such a problem the input space is the richly structured space of possible texts and the output space is the large but largely unstructured list of possible authors. Training data would consist of text author pairs for texts known to have been authored by authors in the list of possible authors.

To solve such a problem one would need to choose a set of features to extract from the underlying documents. Then, using those features and the training data, a classification function would be learned. In this case one might choose a multi-class logistic or probit regression function and learn parameters for it from the labeled training instances. In this chapter we will focus on models that use mixed-membership representations of bag-of-word data as the input features to joint models that can be used, via conditioning, for classification or regression.

The models that we propose in this chapter generalize the author classification example in that the label spaces we will consider are also richly structured (in addition to the input space) and the regression function we learn maps from from the input space to labels in a structured, typically hierarchical label space. In the same way that mixed membership models are surprisingly effective, the resulting hierarchical classification models seem to be too.

1.1.1 Background

Mixed membership models, particularly latent Dirichlet allocation (LDA) [8], 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 each bag-of-words observation consisting of mixing proportions over topic vectors. Each topic is a frequency histogram of terms. Each document is a mixture of topics. This mixture of topics vector is the feature vector we exploit for purposes of document classification.

The idea for using this representation for classification was first considered in a paper on so-calling “supervised” latent Dirichlet allocation (SLDA). SLDA built on LDA by incorporating “supervision” in the form of an observed exponential family response variable per document. Another way of saying this is that SLDA was the first document classifier for documents in the bag-of-words representation that exploited the LDA topic feature representation. SLDA and the subject of this chapter, a generalization of it called hierarchically supervised LDA (HSLDA) [25] are not, strictly speaking, regression models. This is because they model the *joint* distribution of labels and bag-of-words data rather than the conditional distribution of the labels given the input. Having the joint distribution allows us to condition on an input to produce a label for all test instances. Additionally, during training of the model, knowing the labels effects the learning of the low dimensional representation of the input – a characteristic discussed later in this chapter.

The problem of applying structured labels to bag-of-words data can be tackled in a number of different ways. For instance, SLDA can be used to solve this kind of problem directly; however, doing so requires ignoring the hierarchical dependencies amongst the labels. In Section 1.4 we contrast HSLDA with SLDA applied in this way. Other models that incorporate LDA and supervision and so could be used to solve this problem include LabeledLDA [26] and DiscLDA [20]. Various applications of these models to computer vision and document networks have been explored [30, 11]. None of these models, however, leverage dependency structure in the label space.

In other non-LDA-based related work, researchers have classified documents into a hierarchy (a closely related task) using naive Bayes classifiers and support vector machines. Most of this work has been demonstrated on relatively small datasets, small label spaces, and has focused on single label classification without a model of documents such as LDA [23, 13, 19, 10]. The HSLDA model we cover here is not limited in this way.

1.1.2 Intuition

Consider web retail data. Web retailers often have both a browse-able 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 of the free-text product descriptions. One might assume that the free-text descriptions of all of the products in a particular node in the product hierarchy are related. 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. What is more, if topic proportions are indicative of the text describing products that are grouped together, it should be possible to use those proportions to decide (via 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.

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 for instance; more fully 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 [1], product descriptions and catalogs, (e.g. [4] as available from [2]) 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 that allows one to categorize new text documents automatically, suggest labels that might be inaccurate, compute improved similarities between documents for information retrieval purposes, and more.

In this chapter we focus on "is-a" hierarchically structured label spaces, but label spaces that are structured differently may be accommodated with relatively minor modifications. We show 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.

The remainder of this chapter is arranged as follows. Section 1.2 introduces hierarchically supervised LDA (HSLDA), while Section 1.3 details a sampling approach to inference in HSLDA. Section 1.1 reviews related work, and Section 1.4 shows results from applying HSLDA to health care and web retail data.

1.2 Hierarchical Supervised LDA (HSLDA)

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We call individual groups of bag-of-word data documents (although they could be any data represented by bags of features). 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 $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label $l \in \mathcal{L}$, except the root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ 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. 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 is applied to document d or not. In most cases $y_{i,d}$ 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 docu-

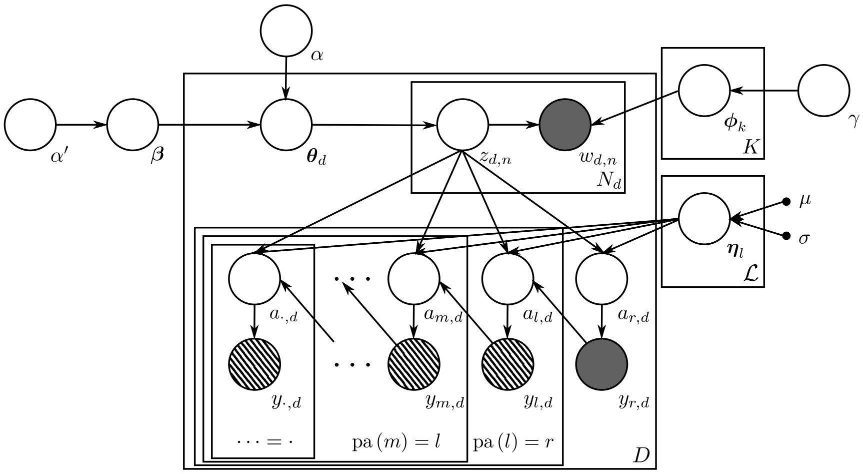

α
α'
$\boldsymbol{\beta}$
$\boldsymbol{\theta}_d$
$z_{d,n}$
$w_{d,n}$
N_d
$\boldsymbol{\phi}_k$
K
γ
μ
σ
$\boldsymbol{\eta}_l$
$\mathcal{L}$
$a_{\cdot,d}$
$a_{m,d}$
$a_{l,d}$
$a_{r,d}$
$y_{\cdot,d}$
$y_{m,d}$
$y_{l,d}$
$y_{r,d}$
$\cdots = \cdot$
$\mathrm{pa}\,(m) = l$
$\mathrm{pa}\,(l) = r$
D

FIGURE 1.1
Hierarchically supervised latent Dirichlet allocation (HSLDA) graphical model.

ment then no descendant of that label may be applied to the same. We assume that at least one label is applied to every document. This is illustrated in Figure 1.1 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).

In HSLDA, the bag-of-word document data is modeled using LDA with full, hierarchical topic estimation (i.e. with the global topic estimated too rather than assumed). Label responses are modeled using a conditional hierarchy of probit regressors. The HSLDA graphical model is given in Figure 1.1.

In the corresponding mathematical generative description of HSLDA to follow, 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.

1.2.1 Ancestral sampling for HSLDA

The following procedure describes how to generate from the HSLDA generative model.

HSLDA Ancestral Sampler

```
1. For each topic $k = 1, \ldots, K$
    •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 $d = 1, \ldots, D$
    •Draw topic proportions $\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \alpha \sim \mathrm{Dir}_K(\alpha \boldsymbol{\beta})$
    •For $n = 1, \ldots, N_d$
        –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
            $a_{l,d} \mid \bar{\mathbf{z}}_d, \boldsymbol{\eta}_l, y_{\mathrm{pa}(l),d} \sim \begin{cases} \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1), & y_{\mathrm{pa}(l),d} = 1 \\ \mathcal{N}(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l, 1)\mathbb{I}(a_{l,d} < 0), & y_{\mathrm{pa}(l),d} = -1 \end{cases}$    (1.1)
        –Apply label $l$ to document $d$ according to $a_{l,d}$
            $y_{l,d} \mid a_{l,d} = \begin{cases} 1 & \text{if } a_{l,d} > 0 \\ -1 & \text{otherwise} \end{cases}$    (1.2)
```

Here $\bar{\mathbf{z}}_d^T = [\bar{z}_1, \ldots, \bar{z}_k, \ldots, \bar{z}_K]$ 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, $\bar{z}_k = N_d^{-1} \sum_{n=1}^{N_d} \mathbb{I}(z_{n,d} = k)$.

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. Note that $y_{l,d}$ can only be applied to document d (set to 1) if its parent label pa(l)

was also applied (these expressions are specific to is-a constraints but could be modified to accommodate different constraints between labels). The regression coefficients $\boldsymbol{\eta}_l$ are independent a priori, however, the hierarchical coupling in this model induces a posteriori dependence. The net effect of this conditional hierarchy of profit regressors is that label predictors deeper in the label hierarchy are able to focus on finding features that distinguish the members of that set, conditioned on the fact that the members are in the same set. It should be clear that one can trivially restrict this hierarchy to a depth one hierarchy. 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. In particular, choosing probit-style auxiliary variable distributions for the $a_{l,d}$'s yields conditional posterior distributions for both the auxiliary variables (1.5) and the regression coefficients (1.4) which are analytic. This simplifies posterior inference substantially. A review of probit regression can be found near the end of this chapter in Section 1.6.3.

In the common case where no negative labels are observed (like the example applications we consider in Section 1.4), the model must be explicitly biased towards generating negative labels in order to keep it from learning to assign positive labels to all documents. This is a common problem in modeling unbalanced data. 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 above. 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 1.4. 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 (1.1) and (1.2). This technique can be used to vary out-of-sample label bias after learning.

1.3 Inference

Our inference goal is to obtain a representation of the posterior distribution of the latent variables in the model. 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 a 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. A very brief review of Markov chain Monte Carlo can be found in Section 1.6.1 at the end of this chapter.

The HSLDA sampler, like the collapsed Gibbs samplers for LDA [18], is itself a collapsed Gibbs sampler in which all of the latent variables that can be analytically marginalized are. Among others, the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$ are analytically marginalized prior to deriving the following conditional distributions for sampling. It will often be the case that the set of labels $\mathcal{L}$ is not fully observed for every document. We will define $\mathcal{L}_d$ to be the subset of labels which have been observed for document d. In an is-a hierarchical regression it is straightforward to marginalize the variables $a_{l',d}$ and $y_{l',d}$ for $l' \in \mathcal{L} \backslash \mathcal{L}_d$ simply by ignoring them. The remaining latent variables (those that are not collapsed out) are $\mathbf{z} = \{z_{1:N_d,d}\}_{d=1,\ldots,D}, \boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{l \in \mathcal{L}}, \mathbf{a} = \{a_{l',d}\}_{l' \in \mathcal{L}_d, d=1,\ldots,D}, \boldsymbol{\beta}, \alpha, \alpha'$ and γ.

1.3.1 Gibbs Sampler

Let $\mathbf{a}$ be the set of all 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. We use the notation $\mathbf{z}_{-(n,d)}$ to denote $\mathbf{z}_d \backslash z_{n,d}$.

First we consider the conditional distribution of $z_{n,d}$ (the assignment variable for each word $n = 1, \ldots, N_d$ in documents $d = 1, \ldots, D$). Following the factorization of the model (refer again to Figure 1.1), 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 [18] we find

$$\begin{aligned} & p\left(z_{n,d} = k \mid \mathbf{z} \backslash z_{n,d}, \mathbf{a}, \mathbf{w}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, \gamma\right) \propto \\ & \quad \left(c^{k,-(n,d)}_{(\cdot),d} + \alpha \boldsymbol{\beta}_k\right) \frac{c^{k,-(n,d)}_{w_{n,d},(\cdot)} + \gamma}{\left(c^{k,-(n,d)}_{(\cdot),(\cdot)} + V\gamma\right)} \prod_{l \in \mathcal{L}_d} \exp\left\{-\frac{\left(\bar{\mathbf{z}}_d^T \boldsymbol{\eta}_l - a_{l,d}\right)^2}{2}\right\} \end{aligned} \tag{1.3}$$

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 $(\cdot)$'s indicate 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 (1.3) by enumeration and implicit normalization.

The conditional posterior distribution of the regression coefficients is given by

$$p(\boldsymbol{\eta}_l \mid \mathbf{z}, \mathbf{a}, \sigma) = \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}) \tag{1.4}$$

$\boldsymbol{\eta}_l$ 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 [5]; the specific form of the update is a standard result from Bayesian normal linear regression [16]. It also is a standard probit regression result that the conditional posterior distribution of $a_{l,d}$ is a truncated normal distribution [5].

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

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 [29]. Sampling $\boldsymbol{\beta}$, the vector of global topic proportions, can be done using the "direct assignment" method of [28]

$$\boldsymbol{\beta} \mid \mathbf{z}, \alpha', \alpha \sim \mathrm{Dir}\left(m_{(\cdot),1} + \alpha', m_{(\cdot),2} + \alpha', \ldots, m_{(\cdot),K} + \alpha'.\right) \tag{1.5}$$

Here $m_{d,k}$ are auxiliary variables that are required to sample the posterior distribution of $\boldsymbol{\beta}$. Their conditional posterior distribution is sampled according to

$$p\left(m_{d,k} = m \mid \mathbf{z}, \mathbf{m}_{-(d,k)}, \boldsymbol{\beta}\right) = \frac{\Gamma\left(\alpha \boldsymbol{\beta}_k\right)}{\Gamma\left(\alpha \boldsymbol{\beta}_k + c^k_{(\cdot),d}\right)} s\left(c^k_{(\cdot),d}, m\right) \left(\alpha \boldsymbol{\beta}_k\right)^m \tag{1.6}$$

where $s\left(n, m\right)$ represents stirling numbers of the first kind.

The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

1.4 Example Applications

Having described HSLDA and given details on how to do inference in it, we turn to demonstrations of using HSLDA to solve problems from two different domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions.

1.4.1 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 US, 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

ADLER Dischage summary text description snippet

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 ADLER**??**. It is worth noting that the coding can be noisy. Human coders sometimes disagree [3], tend to be more specific than sensitive in their assignments [6], and sometimes make mistakes [15].

An automated process would ideally produce a more complete and accurate diagnosis lists. The 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 rules to applications of online rule learning approaches [12, 17, 14]. 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 [21, 27, 24, 22].

The specific dataset we report results for in this chapter was gathered from the NewYork-Presbyterian Hospital clinical data warehouse. It consists of 6,000 discharge summaries and their associated ICD-9 codes (7,298 distinct codes overall), representing all 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 5,000 discharge summaries for training and 1,000 for testing.

---

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

The text of the discharge summaries was tokenized with NLTK.[2] A fixed vocabulary was formed by taking the top 10,000 tokens with 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.

1.4.2 Product Descriptions and Catalogs

Many web-retails 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. Their product categorization data is available as part of the Stanford Network Analysis Platform (SNAP) dataset [2]. A representative sub-tree of the amazon.com DVD produce categorization tree is shown in Figure ADLER**??**. Product descriptions were obtained separately from the Amazon.com website directly. Once such description is

ADLER DVD description

We study the collection of DVDs in the product catalog specifically. The resulting dataset contains 15,130 product descriptions for training and 1,000 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 hierarchal product catalog.

1.4.3 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 tree-conditional regressions (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.

---

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

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 [7]. This comparison model doesn't 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 for the regression coefficients. 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 ($\mu \in \{-3, -2.8, -2.6, \ldots, 1\}$).

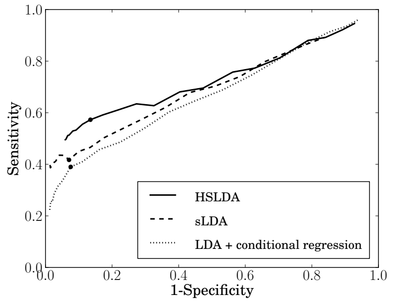
The number of topics for all models was set to $K = 50$, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were all chosen to be gamma with a shape parameter of 1 and a scale parameter of 1000.

1.4.4 Evaluation and Results

We are particularly interested in 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), 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 ances-

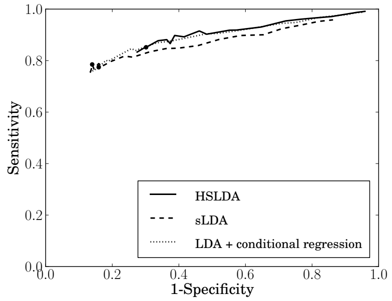
[Clinical data performance.]



[Retail

(a)

(b)



product performance.]

FIGURE 1.2
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.

tors 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 [6]. 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

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

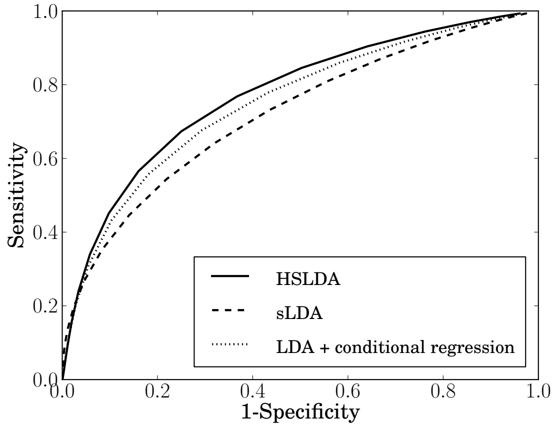


FIGURE 1.3
ROC curve for out-of-sample ICD-9 code prediction varying auxiliary variable threshold. $\mu = -1.0$ for all three models in this figure.

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 1.4.3 demonstrates the performance of the model on the clinical data as an 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 1.4.3.

These results indicate that the full HSLDA model predicts more of the the correct labels at a cost of an increase in the number of false positives relative to the comparison models.

Example topics (as word lists) learned for the discharge data are given below. These word lists are computed by ADLER

ADLER

ADLER could you write a sentence about how these are cleaner than the SLDA or LDA topics?

Figure 1.4.3 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 1.4.3.

Example topics (as word lists) learned for the amazon.com data are given below. T

ADLER

Figure 1.3 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.

1.5 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 that kind 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.

Extensions to this work include unbounded topic cardinality variants 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.

1.6 Appendix

1.6.1 MCMC

1.6.2 Auxiliary Variables

1.6.3 Probit Regression

For reasons that are somewhat obscure to me, statisticians tend to use probit regression for binary classification whereas machine learners tend to use logistic regression. In a recent paper, my collaborators and I found it useful for computational purposes to use probit regression. One can find many good primers on probit regression around the web, but, as we all know, there is almost always space for another.

Figure 1.5 is the cumulative distribution function (cdf) of a $N(0,1)$ distribution. The "probit" function is the inverse of the normal cdf. The normal cdf function, denoted $\Phi(x;\mu,\sigma^2)$ with μ the mean, σ^2 the variance and x the argument, is actually much more relevant in this context.

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 model (a "probit regression model") of the form

$$P(y_i = 1) = \Phi(x_i^T\beta; 0, \sigma^2). \tag{1.7}$$

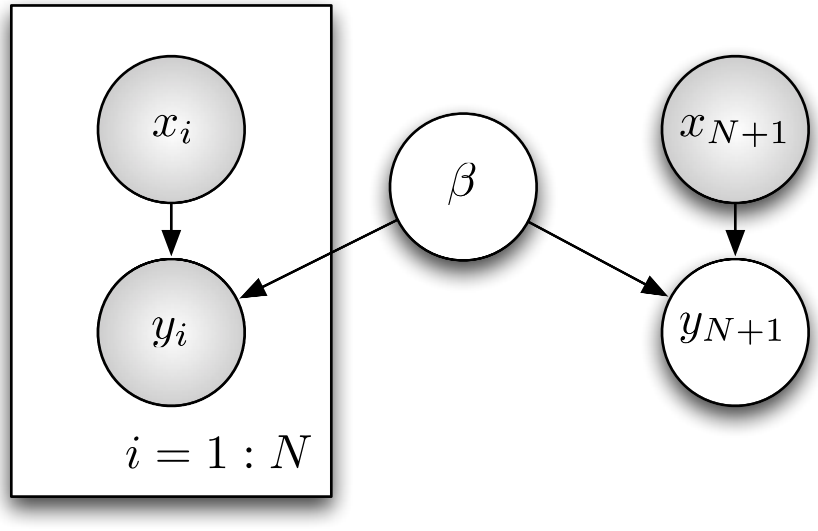

x_i
y_i
$i = 1 : N$
β
x_{N+1}
y_{N+1}

FIGURE 1.4
Probit model, no auxiliary variables

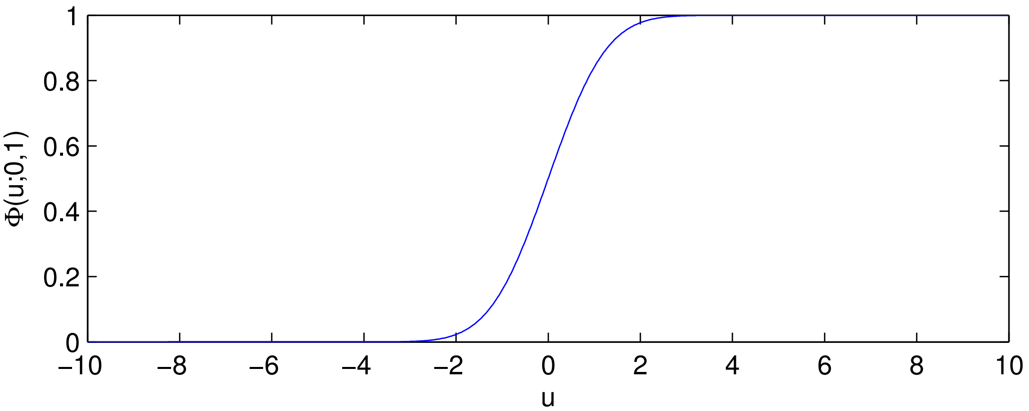


FIGURE 1.5
CDF of $N(0, 1)$

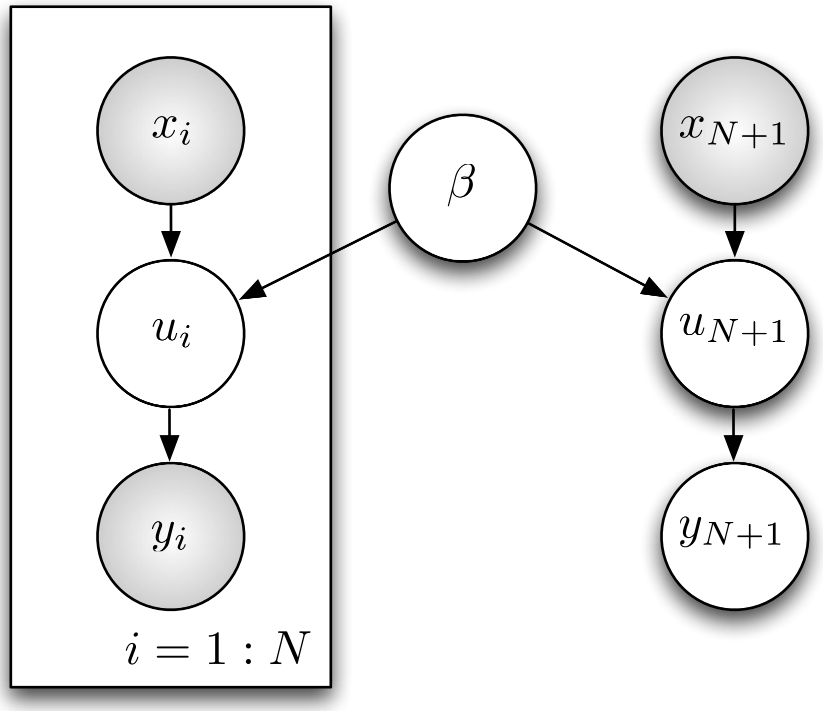

x_i
u_i
y_i
$i = 1 : N$
β
x_{N+1}
u_{N+1}
y_{N+1}

FIGURE 1.6
Probit model

Depending on convention (i.e. binary y_i represented as $\{1, 0\}$ or $\{1, -1\}$) then 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. As always, the close relationship between regression and classification are in full display here : probit regression is a "generalized linear *regression* model" and a "binary classifier."

Figure 1.4 shows the graphical model for probit regression (minus any priors on the regression weights β). 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 $y_{N+1}|x_{N+1}, \beta$. We will be being Bayesian about our inference so here, what we really mean is, 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 $\beta|\{x_i, y_i\}_{i=1}^N$. This brings us to Figure 1.6 which introduces a set of auxiliary variables $\{u_i\}_{i=1}^N$. In this note we will demonstrate that the model in Figure 1.6 is the same as the model in Figure 1.4 when the u_i's are marginalized out and will suggest that inference in the former is computationally easier.

First – what is an auxiliary variable? It is a variable 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, let's forget about the i index for a second 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 write down the joint distribution of these quantities (according to the graphical model that includes auxiliary variables).

$$P(y, x, u) = P(y|u)P(u|x, \beta) \tag{1.8}$$

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

$$P(y, x) = \int P(y, x, u)du = \int P(y|u)P(u|x, \beta)du. \tag{1.9}$$

If by MCMC 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 what $P(u|x, \beta)$ are. Let's try $y = \text{sign}(u)$ and $u \sim N(x^T\beta, \sigma^2)$. These choices are nice in a particular way. First let's verify that the marginalization of u out of this model results in the model specification in Equation 1.7.

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

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).

OK. So, now, we have 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 and 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. Sampling β from its posterior distribution typically is quite simple; certainly more so that sampling β without the u auxiliary variables.

Bibliography

[1] DMOZ open directory project. `http://www.dmoz.org/`, 2002.

[2] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[3] The computational medicine center's 2007 medical natural language processing challenge. http://www.computationalmedicine.org/challenge/previous, 2007.

[4] Amazon, Inc. `http://www.amazon.com/`, 2011.

[5] J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669, 1993.

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

[7] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20:121–128, 2008.

[8] D. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[9] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.

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

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

[12] K. Crammer, M. Dredze, K. Ganchev, P.P. Talukdar, and S. Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[13] S. Dumais and H. Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 256–263, New York, NY, USA, 2000. ACM.

[14] R. Farkas and G. Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[15] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi. 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, 2010.

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

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

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[19] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[20] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[21] L. Larkey and B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, University of Massachussets, 1995.

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

[23] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[24] S. Pakhomov, J. Buntrock, and C. Chute. 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(5):516–525, 2006.

[25] A. Perotte, N. Bartlett Noıemie Elhadad, and F. Wood. Hierarchically supervised latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, to appear, 2012.

[26] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. 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*, pages 248–256, 2009.

[27] B. Ribeiro-Neto, A. Laender, and L. De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401, 2001.

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[29] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[30] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, pages 1903–1910, 2009.

Index

basis, 9
    ordered, 9

linear combination, 8
    nontrivial, 8
    trivial, 8
linearly independent, 9
    maximal, 9

matrix
    scalar multiple of, 9
    scalar multiplication, 9

spanning set, 8
    minimal, 9
subspace, 7

vector space
    subspace of, 7