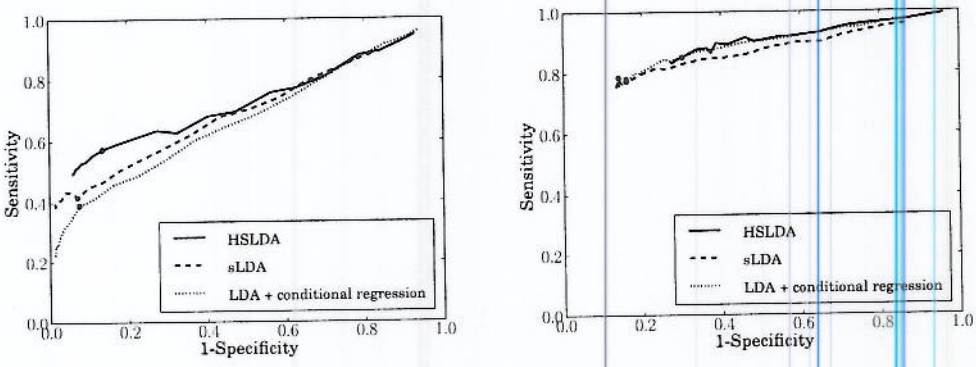

(a) Clinical data performance.

(b) Retail product performance.

FIGURE 14.4
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.

14.4.4 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), 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 η_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 pos-

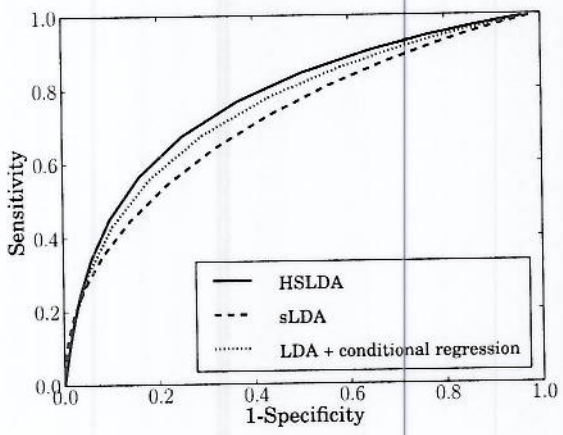

FIGURE 14.5
ROC curve for out-of-sample ICD-9 code prediction varying auxiliary variable threshold. $\mu = -1.0$ for all three models in this figure.

itive 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 14.4(a) 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 14.4(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 14.4(a), HSLDA outperforms no worse than the comparison models across the full range of specificities.

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 14.4(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 14.4(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 14.5 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.

14.5 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 14.4, 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 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).

14.6 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 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 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 (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 car-

dinality 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.

14.7 Appendix

14.7.1 Probit Regression

The entire section

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

$$P(y_i = 1) = \Phi(x_i^T \beta; 0, \sigma^2). \tag{14.7}$$

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 this efficiently one can introduce a set of auxiliary variables $\{u_i\}_{i=1}^N$.

By auxiliary variable 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

$$P(y, x, u) = P(y|u)P(u|x, \beta). \tag{14.8}$$

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

$$P(y, x) = \int P(y, x, u)du = \int P(y|u)P(u|x, \beta)du. \tag{14.9}$$

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 (14.7).

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

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 us "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; 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 Iinformation 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., 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.

Ramage, D., Hall, D., Nallapati, R., 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., 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.