

Discriminative Mixed-membership Models

Hanhuai Shan, Arindam Banerjee
Dept of Computer Science & Engineering
University of Minnesota, Twin Cities
{shan,banerjee}@cs.umn.edu

Nikunj C. Oza
Intelligent Systems Division
NASA Ames Research Center
nikunj.c.oza@nasa.gov

Abstract

Although mixed-membership models have achieved great success in unsupervised learning, they have not been widely applied to classification problems. In this paper, we propose a family of discriminative mixed-membership models for classification by combining unsupervised mixed-membership models with multi-class logistic regression. In particular, we propose two variants respectively applicable to text classification based on latent Dirichlet allocation and usual feature vector classification based on mixed-membership naive Bayes models. The proposed models allow the number of components in the mixed membership to be different from the number of classes. We propose two variational inference based algorithms for learning the models, including a fast variational inference which is substantially more efficient than mean-field variational approximation. Through extensive experiments on UCI and text classification benchmark datasets, we show that the models are competitive with the state of the art, and can discover components not explicitly captured by the class labels.

1 Introduction

In recent years, mixed-membership (MM) models have been found wide application in a variety of domains, such as topic modeling [6], bioinformatics [1] and social network analysis [12]. A key advantage of such models is that they provide a succinct and interpretable representation of otherwise large and high-dimensional datasets. However, one important restriction of most existing MM models is that they are unsupervised models and cannot leverage class label information for classification. On the other hand, while popular classification algorithms, such as support vector machines (SVM) [7] and logistic regression (LR) [17], perform well on classification, the classifier itself is often hard to interpret. The above observation motivates our current work on designing accurate discriminative classification algorithms while leveraging mixed-membership models for

interpretability.

Supervised latent Dirichlet allocation (SLDA) [5] is such a mixed-membership model which takes response variables into account. However, the response variables in SLDA are real numbers assumed to be generated from a normal linear model, which is different from categorical labels in the context of classification. In principle, the authors proposed a general framework to extend SLDA to deal with other types of response variables, including categorical labels, based on generalized linear models (GLM) [14]. However, efficient inference in the general case is difficult without the good properties of Gaussian distribution. In addition, SLDA is only designed to handle text data or a sequence of homogeneous tokens, while several real world classification problems involve heterogeneous features with measured values, e.g., most datasets in the UCI benchmark.

In this paper, we propose discriminative¹ mixed-membership (DM) models by combining multi-class logistic regression with unsupervised MM models. In particular, we consider two variants—discriminative latent Dirichlet allocation (DLDA) and discriminative mixed-membership naive Bayes (DMNB). DLDA is applicable to text classification and uses latent Dirichlet allocation (LDA) [6] as the underlying MM model. DMNB is applicable to non-text classification involving numerical feature vectors and uses mixed-membership naive Bayes (MNB) [2] as the underlying MM model. The mixed-membership representation generated by DM is biased by class labels and can be viewed as a supervised dimensionality reduction. Further, since DM allows the number of components k in the mixed membership to be different from the number of classes c , the model often discovers additional latent structure beyond what implies by the class labels with a larger k . To learn the model, we propose two families of variational inference algorithms: one is based on ideas originally proposed in [6] and the other is more efficient in space and time complexity by using a significantly less number of parameters. Unlike Taylor expansion based approximations suggested in [5],

¹“Discriminative” here does not mean a discriminative model, but a generative model used for classification instead of clustering.

the proposed inference algorithms maintain the lower bound maximization strategy used in variational inference.

Recently, there has been an increasing interest in mixed-membership models combining supervision information. Other than SLDA, [10] proposed labeled latent Dirichlet allocation to incorporate functional annotation of known genes to guide gene clustering. [13] proposed DiscLDA which determines document position on topic simplex with guidance of labels. [15] proposed a Dirichlet-multinomial regression which accommodates different types of meta-data, including labels. [19] proposed a correlated labeling model for multi-label classification. [18] extends SLDA for image classification and annotation.

The rest of the paper is organized as follows: In Section 2, we give a brief overview of mixed-membership models. In Section 3, we propose discriminative mixed-membership models. In Section 4, a variational approach for learning DM is given. We present the experimental results in Section 5 and conclude in Section 6. In the following sections, mixed-membership models particularly refer to LDA and MNB.

2 Generative mixture models

In this section, we give an overview on two mixed-membership models—latent Dirichlet allocation and mixed-membership naive Bayes models. We also briefly introduce supervised latent Dirichlet allocation.

2.1 Latent Dirichlet allocation

LDA [6] is a three-level Bayesian model as an extension of finite mixture models (FMM) for topic modeling. Instead of having a fixed component proportion π for all data points as in FMM, LDA maintains a separate component proportion π over k components for each document $x_{1:N}$, and π is sampled from a Dirichlet distribution $\text{Dir}(\alpha)$. For a sequence of words in a document $x_{1:N}$ and the corresponding sequence of components (topics) $z_{1:N}$, LDA has a density of the form

$$p(x_{1:N}|\alpha, \beta_{1:k}) = \int_{\pi} p(\pi|\alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n|\pi) p(x_n|z_n, \beta_{1:k}) \right) d\pi,$$

where $\beta_{1:k} = \{\beta_i, [i]_1^k\}$ ($[i]_1^k \equiv i = 1, \dots, k$) is a collection of parameters for k component distributions, each of which is a Discrete distribution over all words in the dictionary.

Getting a closed form expression for the marginal density $p(x_{1:N}|\alpha, \beta_{1:k})$ is intractable. Variational inference [6] and Gibbs sampling [11] are two most popular approaches proposed to address the problem.

2.2 Supervised LDA

Supervised latent Dirichlet allocation (SLDA) [5] is an extension of LDA which accommodates the response variables other than the documents. The response variable is assumed to be generated from a normal linear model $N(\eta^T \bar{z}, \sigma^2)$, where η and σ^2 are the parameters and the covariates $\bar{z} = \sum_{n=1}^N z_n/N$ are the empirical average frequencies of each latent topic in the document. The density function of SLDA is given as follows:

$$p(x_{1:N}, y|\alpha, \beta_{1:k}, \eta, \sigma^2) = \int_{\pi} p(\pi|\alpha) \sum_{z_{1:N}} \left(\prod_{n=1}^N p(z_n|\pi) p(x_n|z_n, \beta_{1:k}) \right) p(y|z_{1:N}, \eta, \sigma^2) d\pi.$$

Since y is assumed to be generated from a univariate normal linear model, SLDA is constrained to deal with one dimensional real-valued response variables.

2.3 Mixed-membership naive Bayes

Although LDA achieves a good performance in topic modeling, it suffers from two limitations [2]: (1) LDA cannot deal with data points with measured feature values. (2) LDA cannot deal with data points with heterogeneous features. MNB relaxes these limitations by introducing a separate exponential family distribution [3] for each feature. It is designed to deal with sparse and heterogeneous feature vectors. Given a data point $x_{1:N}$, the density function of MNB model with k components is as follows:

$$p(x_{1:N}|\alpha, \Theta) = \int_{\pi} p(\pi|\alpha) \left(\prod_{\substack{n=1 \\ \exists x_n}}^N \sum_{z_n} p(z_n|\pi) p_{\psi_n}(x_n|z_n, \theta_n) \right) d\pi,$$

where $\exists x_n$ indicates that the model only considers the non-missing features, $\Theta = \{\theta_n, [n]_1^N\}$ are the parameters for the distributions of N features respectively, and each $\theta_n = \{\theta_{ni}, [i]_1^k\}$ are the parameters over k components of feature n . $p_{\psi_n}(x_n|z_n, \theta_n)$ is an exponential family distribution with a form of $p_{\psi}(x|\theta) = \exp(\langle x, \theta \rangle - \psi(\theta)) p_0(x)$, where θ is the natural parameter, $\psi(\cdot)$ is the cumulant function, and $p_0(x)$ is a non-negative base measure. ψ determines a particular family, such as Gaussian, Poisson, etc., and θ determines a particular distribution in that family.

3 Discriminative mixed-membership models

We motivate discriminative mixed-membership models by considering two important limitations of SLDA [5] which prevent it from being used as a discriminative classification model: First, the response variables in SLDA are

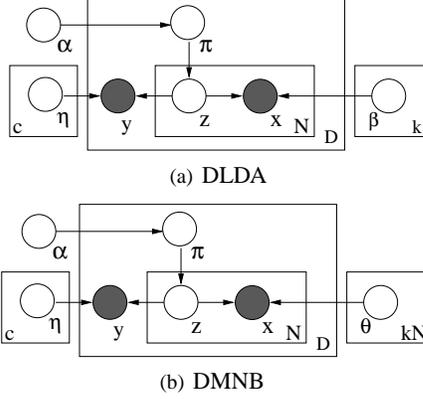


Figure 1. Graphical models for DLDA and DMNB.

univariate real numbers assumed to be generated from a normal linear model, whereas the response variables, i.e., labels, are discrete categories in the classification setting. Although the authors pointed out that the response variables can be of various types obtained from generalized linear models, variational inference is difficult in the general case. While a Taylor expansion is recommended [5] to obtain an approximation of the log-likelihood, such an approach forgoes the lower bound guarantee of variational inference. Second, like LDA, SLDA is designed for text data viewed as a sequence of homogeneous tokens. However, most non-text classification tasks, e.g., the UCI benchmark datasets, have features with measured values. Further, the features could be heterogeneous with different semantics, different ranges of values, etc., such as a customer’s age, occupation and zip code. SLDA is not designed for such data.

The proposed family of discriminative mixed membership models overcome both limitations. In particular, DLDA is a variant of SLDA which accommodates categorical response variables and is hence suitable for text classification tasks. Further, DMNB is a variant of MNB, i.e., a generalization suitable for discriminative classification with (non-text) heterogeneous feature vectors. In principle, DMNB works for sparse data, but the sequel only considers the non-sparse case for ease of exposition.

3.1 Discriminative LDA

Assuming there are c classes and the number of components we choose is k , the graphical model for DLDA is given in Figure 1(a), where α is a k -dimensional parameter of a Dirichlet distribution, $\beta_{1:k}$ are the parameters for k component distributions over the words with each component referring to a topic, and $\eta_{1:c} = [\eta_1, \dots, \eta_c]^T$ is a matrix with c k -dimensional logistic regression parameters as the rows, where η_c is a zero vector by default, so we only use $\eta_{1:c-1}$ as the parameter to be estimated. The generative process for each document $x_{1:N}$ is given as follows:

1. Choose a component proportion $\pi \sim \text{Dirichlet}(\alpha)$.

2. For each word in the document,
 - (a) Choose a component $z_n = i \sim \text{Discrete}(\pi)$.
 - (b) Choose a word $x_n \sim \text{Discrete}(\beta_i)$.
3. Choose the label from a multi-class logistic regression
$$y \sim \text{LR} \left(\frac{\exp(\eta_h^T \bar{z})}{1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})} \right), [h]_1^{c-1}.$$

\bar{z} is an average of $z_{1:N}$ over all observed words, where each z_n is a k -dimensional unit vector with only the i^{th} entry being 1 if it denotes the i^{th} component. The categorical response variable y can be considered as a sample generated from the Discrete distribution $(p_1, \dots, p_{c-1}, 1 - \sum_{h=1}^{c-1} p_h)$ where $p_h = \frac{\exp(\eta_h^T \bar{z})}{1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})}$ for $[h]_1^{c-1}$. In two-class classification, y is 0 or 1 generated from $\text{Bernoulli}(\frac{1}{1 + \exp(-\eta^T \bar{z})})$, i.e., the model needs only one η in the two-class case.

There are two important properties of DLDA and DMNB models in general: (1) The k -dimensional mixed membership \bar{z} effectively serves as a low dimensional representation of the original document. While \bar{z} in LDA is inferred in an unsupervised way, it is obtained from a supervised dimensionality reduction in DLDA. We give the explanation in Section 4. (2) DLDA allows the number of classes c and the number of components k in the generative model to be different. If k was forced to be equal to c , for problems with a small number of classes, \bar{z} would have been a rather coarse representation of the document. In particular, for two-class problems, \bar{z} would lie on the 2-simplex which may not be an informative representation for classification purposes. Decoupling the choice of k from c prevents such pathologies. In principle, we may find a proper k using Dirichlet process mixture models [4].

From the generative model, the joint distribution of latent and observable variables for DLDA is given by

$$p(\pi, z_{1:N}, x_{1:N}, y | \alpha, \beta_{1:k}, \eta_{1:c-1}) \quad (1)$$

$$= p(\pi | \alpha) \left(\prod_{n=1}^N p(z_n | \pi) p(x_n | z_n, \beta_{1:k}) \right) p(y | z_{1:N}, \eta_{1:c-1}).$$

Integrating (1) over π and summing it over $z_{1:N}$ yields the marginal distribution of $(x_{1:N}, y)$:

$$p(x_{1:N}, y | \alpha, \beta_{1:k}, \eta_{1:c-1}) = \int_{\pi} p(\pi | \alpha) \quad (2)$$

$$\sum_{z_{1:N}} \left(\prod_{n=1}^N p(z_n | \pi) p(x_n | z_n, \beta_{1:k}) \right) p(y | z_{1:N}, \eta_{1:c-1}) d\pi.$$

The probability of the entire data set of D documents and their labels $(\mathcal{X} = \{x_d, [d]_1^D\}, \mathcal{Y} = \{y_d, [d]_1^D\})$ is given by

$$p(\mathcal{X}, \mathcal{Y} | \alpha, \beta_{1:k}, \eta_{1:c-1}) = \prod_{d=1}^D \int_{\pi_d} p(\pi_d | \alpha) \quad (3)$$

$$\sum_{z_{d,1:N}} \left(\prod_{n=1}^N p(z_{dn} | \pi_d) p(x_{dn} | z_{dn}, \beta_{1:k}) \right) p(y_d | z_{d,1:N}, \eta_{1:c-1}) d\pi_d.$$

(a) ϕ	
Std DLDA	$\phi_{ni} \propto \exp\left(\Psi(\gamma_i) - \Psi(\sum_{l=1}^k \gamma_l) + \sum_{v=1}^V \mathbb{1}(x_n \sim v) \log \beta_{iv} + \frac{1}{N} \sum_{h=1}^{c-1} (\eta_{hi} y_h - \exp(\eta_{hi})/\xi)\right)$
Fast DLDA	$\phi_i \propto \exp\left(\Psi(\gamma_i) - \Psi(\sum_{l=1}^k \gamma_l) + \frac{1}{N} \sum_{n=1}^N \sum_{v=1}^V \mathbb{1}(x_n \sim v) \log \beta_{iv} + \frac{1}{N} \sum_{h=1}^{c-1} (\eta_{hi} y_h - \exp(\eta_{hi})/\xi)\right)$
Std DMNB	$\phi_{ni} \propto \exp\left(\Psi(\gamma_i) - \Psi(\sum_{l=1}^k \gamma_l) + \left(-\frac{(x_n - \mu_{ni})^2}{2\sigma_{ni}^2} - \log \sqrt{2\pi\sigma_{ni}^2}\right) + \frac{1}{N} \sum_{h=1}^{c-1} (\eta_{hi} y_h - \exp(\eta_{hi})/\xi)\right)$
Fast DMNB	$\phi_i \propto \exp\left(\Psi(\gamma_i) - \Psi(\sum_{l=1}^k \gamma_l) + \frac{1}{N} \sum_{n=1}^N \left(-\frac{(x_n - \mu_{ni})^2}{2\sigma_{ni}^2} - \log \sqrt{2\pi\sigma_{ni}^2}\right) + \frac{1}{N} \sum_{h=1}^{c-1} (\eta_{hi} y_h - \exp(\eta_{hi})/\xi)\right)$

(b) γ		(c) ξ	
Std DLDA/DMNB	$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni}$	Std DLDA/DMNB	$\xi = 1 + \frac{1}{N} \sum_{h=1}^{c-1} \sum_{i=1}^k \sum_{n=1}^N \phi_{ni} \exp(\eta_{hi})$
Fast DLDA/DMNB	$\gamma_i = \alpha_i + N \phi_i$	Fast DLDA/DMNB	$\xi = 1 + \sum_{h=1}^{c-1} \sum_{i=1}^k \phi_i \exp(\eta_{hi})$

Table 1. Updates for variational parameters.

3.2 Discriminative MNB

Discriminative MNB is similar with DLDA except that it keeps separate distributions for each feature. Given the graphical model in Figure 1(b), the generative process for $x_{1:N}$ is as follows:

1. Choose a component proportion $\pi \sim \text{Dirichlet}(\alpha)$.
2. For each feature in the data point
 - (a) Choose a component $z_n = i \sim \text{Discrete}(\pi)$.
 - (b) Choose a feature value $x_n \sim p_{\psi_n}(x_n | \theta_{ni})$
3. Choose the label from a multi-class logistic regression $y \sim \text{LR}\left(\frac{\exp(\eta_h^T \bar{z})}{1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})}\right), [h]_1^{c-1}$.

$p_{\psi_n}(x_n | \theta_{ni})$ in 2(b) is an exponential family distribution [3]. Comparing DMNB with DLDA, for each component/topic i , DLDA has only one discrete distribution to generate features (words), while DMNB has separate distributions $p_{\psi_n}(x_n | \theta_{ni})$ for different feature n . These component distributions for DMNB could be of different types, or a same type with different parameters, just like in naive Bayes. Therefore, DMNB is more flexible than DLDA to deal with heterogenous features with measured values by choosing a proper distribution for each feature.

DMNB could be considered as a generalization of naive Bayes (NB) classifier extended in the following aspects: First, NB shares a component among all features, but DMNB has a separate component for each feature and maintains a Dirichlet-multinomial prior on all possible combination of component assignments. Therefore the components for different features might be different in DMNB, and NB could be considered as a special case when z_n is the same for all features. Second, NB uses the shared component as a class indicator, whereas DMNB uses the mixed membership over separate components as inputs to a logistic regression model which finally generates the class label. Third, NB requires $k=c$ while DMNB does not. In principle, DMNB could be applied whenever naive Bayes is applicable.

A special case of DMNB is when each feature n is assumed to be generated from one of k Gaussian distributions with the mean $\mu_n = \{\mu_{ni}, [i]_1^k\}$ and the variance $\sigma_n^2 = \{\sigma_{ni}^2, [i]_1^k\}$. The marginal distribution of $(x_{1:N}, y)$ is:

$$p(x_{1:N}, y | \alpha, \mu_{1:N,1:k}, \sigma_{1:N,1:k}^2, \eta_{1:c-1}) = \int_{\pi} p(\pi | \alpha) \sum_{z_{1:N}} \left(\prod_{n=1}^N p(z_n | \pi) p(x_n | z_n, \mu_{n,1:k}, \sigma_{n,1:k}^2) \right) p(y | z_{1:N}, \eta_{1:c-1}) d\pi. \quad (4)$$

The probability of the entire data set $(\mathcal{X} = \{x_d, [d]_1^D\}, \mathcal{Y} = \{y_d, [d]_1^D\})$ is given by

$$p(\mathcal{X}, \mathcal{Y} | \alpha, \mu_{1:N,1:k}, \sigma_{1:N,1:k}^2, \eta_{1:c-1}) = \prod_{d=1}^D \int_{\pi_d} p(\pi_d | \alpha) \sum_{z_{d,1:N}} \left(\prod_{n=1}^N p(z_{dn} | \pi) p(x_{dn} | z_{dn}, \mu_{n,1:k}, \sigma_{n,1:k}^2) \right) p(y_d | z_{d,1:N}, \eta_{1:c-1}) d\pi_d. \quad (5)$$

4 Inference and parameter estimation

Since DM models assume a generative process for both labels as well as the data points, instead of using labels directly to train a classifier, we use both \mathcal{X} and \mathcal{Y} as samples from the generative process to estimate the parameters of DM models such that the likelihood of observing $(\mathcal{X}, \mathcal{Y})$ is maximized. Unlike naive Bayes [9], the parameters cannot be directly estimated from the class labels due to the latent mixed memberships. In particular, due to the latent variables, the computation of the likelihood in (2) and (4) is intractable. In this section, we present two alternative approaches to obtain a variational approximation of the log-likelihood and propose an expectation maximization (EM)-style algorithm to iteratively obtain better estimates of the model parameters. Finally, we show how the estimated parameters can be used to do prediction on test data.

	Iris	Pima	Vowel	Wine	Wpbc	Ecoli	Iono	Sonar	Seg
D	150	768	990	178	198	336	351	208	2310
N	4	8	11	13	34	7	34	60	19
c	3	2	11	3	2	8	2	2	7

Table 2. UCI Data.

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
D	4226	3893	3000	3000	3000
V	604	5923	7666	10083	5932
Classes	passenger	aeronautics	atheism	guns	graphics
	flight crew	medicine	baseball	midwest	windows
	mainten-	information-	space	politics	ms-
	nance	retrieval			windows

Table 3. Text Data.

4.1 Variational approximation

For each data point, to obtain a tractable lower bound to $\log p(x_{1:N}, y | \alpha, \Lambda, \eta_{1:c-1})^2$, we introduce a variational distribution $q(\pi, z_{1:N} | \Omega)^3$ as an approximation of the true posterior distribution $p(\pi, z_{1:N} | \alpha, \Lambda, \eta_{1:c-1})$ over the latent variables, where Ω is the set of variational parameters. By a direct application of Jensen’s inequality [6], the lower bound to $\log p(x_{1:N}, y | \alpha, \Lambda, \eta_{1:c-1})$ is given by:

$$\log p(x_{1:N}, y | \alpha, \Lambda, \eta_{1:c-1}) \geq E_q[\log p(\pi, z_{1:N}, x_{1:N}, y | \alpha, \Lambda, \eta_{1:c-1})] + H(q(\pi, z_{1:N})). \quad (6)$$

We use L to denote the lower bound. Following [6] and noticing that $x_{1:N}$ and y are conditionally independent given $z_{1:N}$, we have

$$\begin{aligned} L = & E_q[\log p(\pi | \alpha)] + E_q[\log p(z_{1:N} | \pi)] \\ & + E_q[\log p(x_{1:N} | z_{1:N}, \Lambda)] - E_q[\log q(\pi)] - E_q[\log q(z_{1:N})] \\ & + E_q[\log p(y | z_{1:N}, \eta_{1:c-1})]. \end{aligned} \quad (7)$$

We propose two different variational distributions $q(\pi, z_{1:N})$. Following [6], we consider

$$q_1(\pi, z_{1:N} | \gamma, \phi_{1:N}) = q_1(\pi | \gamma) \prod_{n=1}^N q_1(z_n | \phi_n), \quad (8)$$

where $q_1(\pi | \gamma)$ is a Dirichlet distribution for π and each $q_1(z_n | \phi_n)$ is a Discrete distribution for z_n . Also, we propose

$$q_2(\pi, z_{1:N} | \gamma, \phi) = q_2(\pi | \gamma) \prod_{n=1}^N q_2(z_n | \phi), \quad (9)$$

where $q_2(\pi | \gamma)$ is a Dirichlet distribution for π and $q_2(z_n | \phi)$ is a Discrete distribution for all z_n . In both q_1 and q_2 , we have a k -dimensional Dirichlet(γ) for each data point, but we have a k -dimensional Discrete(ϕ_n) for each of N features in q_1 and only one Discrete(ϕ) for all features in q_2 . By keeping a substantially smaller number of parameters,

² Λ denotes $\beta_{1:k}$ for DLDA and $\theta_{1:kN}$ for DMNB.

³To avoid clutter, we do not show the free variational parameters of q unless necessary in the sequel.

q_2 is space efficient especially for high-dimensional data. It is also time efficient with a substantially smaller number of parameters to optimize over. q_1 and q_2 determine two different variational inference algorithms. We call the first one “standard variational inference” and the second one “fast variational inference”, which accordingly yield standard DM/MM (Std DM/MM) models as opposed to Fast DM/MM models respectively. In the sequel, we use q to denote q_1 or q_2 unless otherwise necessary.

Given the variational distribution as in (8) or (9), the first five terms in the lower bound (7) can be easily obtained following LDA or MNB depending on which DM model we are using. The most difficult part is the last term, which cannot be computed exactly even after introducing the variational distribution q , so further approximation is needed. We give the expression for the last term here, the details of derivation could be found in the Appendix. For standard DM models, we have

$$\begin{aligned} & E_q[\log p(y | z_{1:N}, \eta_{1:c-1})] \\ & \geq \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^k \phi_{ni} \left(\sum_{h=1}^{c-1} \eta_{hi} y_h - \frac{1}{\xi} \sum_{h=1}^{c-1} \exp(\eta_{hi}) \right) + \left(1 - \frac{1}{\xi} - \log \xi\right), \end{aligned} \quad (10)$$

and for Fast DM models, we have

$$\begin{aligned} & E_q[\log p(y | z_{1:N}, \eta_{1:c-1})] \\ & \geq \sum_{i=1}^k \phi_i \left(\sum_{h=1}^{c-1} \eta_{hi} y_h - \frac{1}{\xi} \sum_{h=1}^{c-1} \exp(\eta_{hi}) \right) + \left(1 - \frac{1}{\xi} - \log \xi\right), \end{aligned} \quad (11)$$

where $\xi > 0$ is a new variational parameter introduced to obtain a lower bound for the last term in (7).

4.1.1 Inference

Given a choice of model parameters $(\alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)})$, the lower bound to the log-likelihood for each data point in (7) becomes a function of the variational parameters $L(\alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)}, \gamma, \phi, \xi)$. The goal of the inference step is to obtain the tightest lower bound to the true log-likelihood, which is achieved by maximizing $L(\alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)}, \gamma, \phi, \xi)$ with respect to γ, ϕ and ξ . The results of the variational parameters are given in Table 1, where $\mathbb{1}(x_n \sim v)$ takes value 1 if v is the index for the n^{th} word of the document in the dictionary and 0 otherwise, and V is total number of the words in the dictionary.

From Table 1, $\bar{\phi} = [\sum_{n=1}^N \phi_{n1}/N, \dots, \sum_{n=1}^N \phi_{nk}/N]$ for DM and $\phi = [\phi_1, \dots, \phi_k]$ for Fast DM actually give the posterior of \bar{z} , i.e., the low-dimension representation of each data point. Note that the last term in all expressions of ϕ contains y , showing that the low-dimension representation not only depends on $x_{1:N}$, but also depends on y , which means DM models achieve supervised dimension reduction. Removing the last term gives the expression of ϕ in the corresponding unsupervised settings.

	Iris	Pima	Vowel	Wine	Wpbc	Ecoli	Iono	Sonar	Seg
Std	0.9200	0.6500	0.4535	0.9706	0.7737	0.7895	0.6829	0.6300	0.6514
MNB	± 0.0613	± 0.0552	± 0.0299	± 0.0500	± 0.0704	± 0.0629	± 0.0579	± 0.0789	± 0.0293
Fast	0.9466	0.6868	0.4969	0.9470	0.7789	0.7950	0.7486	0.6100	0.6333
MNB	± 0.0688	± 0.0486	± 0.0332	± 0.0647	± 0.0692	± 0.0595	± 0.0643	± 0.0516	± 0.0637
Std	0.9466	0.6553	0.6192	0.9647	0.7632	0.7788	0.7314	0.6000	0.6398
DMNB	± 0.0525	± 0.571	± 0.0571	± 0.0411	± 0.0832	± 0.0554	± 0.0895	± 0.0822	± 0.0397
Fast	0.9600	0.6645	0.6596	0.9765	0.7632	0.8060	0.8031	0.6596	0.7632
DMNB	± 0.0644	± 0.0632	± 0.0409	± 0.0324	± 0.0832	± 0.0762	± 0.1291	± 0.0918	± 0.0507

Table 4. Accuracy on UCI with $k = c$.

	Iris	Pima	Vowel	Wine	Wpbc	Ecoli	Iono	Sonar	Seg
Std	1.8850	3.5525	24.8176	2.2563	3.0255	4.6465	5.2036	4.8948	120.2613
DMNB	± 0.2232	± 1.5372	± 8.7263	± 0.2505	± 0.6129	± 1.1336	± 3.1054	± 4.509	± 77.2722
Fast	1.1651	2.0802	16.1795	1.0988	1.5965	3.9679	0.8220	1.0252	25.368
DMNB	± 0.1771	± 0.1668	± 0.4393	± 0.0425	± 0.2124	± 0.3937	± 0.0077	± 0.0751	± 6.3268

Table 5. Running time (seconds) of standard DMNB and Fast DMNB on UCI data with $k = c$.

4.1.2 Parameter estimation

Variational parameters $(\phi^*, \gamma^*, \xi^*)$ from the inference step gives the optimal lower bound to the log-likelihood of each pair of $(x_{1:N}, y)$. Since we cannot maximize $\log p(\mathcal{X}, \mathcal{Y} | \alpha, \Lambda, \eta_{1:c-1})$ directly, we maximize the aggregate lower bound $\sum_{d=1}^D L(\phi_d^*, \gamma_d^*, \xi_d^*, \alpha, \Lambda, \eta_{1:c-1})$ over all data points with respect to α, Λ and $\eta_{1:c-1}$ respectively to obtain the estimated parameters. The estimations of α and Λ are the same as in the corresponding MM models [6, 2]. As for η , we have

$$\eta_{hi} = \log \frac{\sum_{d=1}^D \sum_{n=1}^{N_d} y_{dh} \phi_{dni} / N_d}{\sum_{d=1}^D \sum_{n=1}^{N_d} \phi_{dni} / (N_d \xi_d)}$$

for standard DM models, and

$$\eta_{hi} = \log \frac{\sum_{d=1}^D \phi_{di} y_{dh}}{\sum_{d=1}^D \phi_{di} / \xi_d}$$

for Fast DM models.

4.2 Variational EM algorithm

We propose an EM-style algorithm to find out the optimal model parameters alternatively. Given $(\alpha^{(t-1)}, \Lambda^{(t-1)}, \eta_{1:c-1}^{(t-1)})$ from the initial guess or the last iteration. The algorithm alternates between the following two steps until convergence:

1. E-step: Given $(\alpha^{(t-1)}, \Lambda^{(t-1)}, \eta_{1:c-1}^{(t-1)})$, for each data point, find the variational parameters

$$(\phi_d^{(t)}, \gamma_d^{(t)}, \xi_d^{(t)}) = \operatorname{argmax}_{(\phi_d, \gamma_d, \xi_d)} L(\phi_d, \gamma_d, \xi_d, \alpha^{(t-1)}, \Lambda^{(t-1)}, \eta_{1:c-1}^{(t-1)}),$$

then $L(\phi_d^{(t)}, \gamma_d^{(t)}, \xi_d^{(t)}; \alpha, \Lambda, \eta_{1:c-1})$ gives a lower bound to $\log p(x_d, y_d | \alpha, \Lambda, \eta_{1:c-1})$.

2. M-step: Maximizing the aggregate lower bound yields an improved estimate of model parameters:

$$(\alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)}) = \operatorname{argmax}_{(\alpha, \Lambda, \eta)} \sum_{d=1}^D L(\gamma_d^{(t)}, \phi_d^{(t)}, \xi_d^{(t)}; \alpha, \Lambda, \eta_{1:c-1}).$$

After t iterations, the objective function becomes $\sum_{d=1}^D L(\gamma_d^{(t)}, \phi_d^{(t)}, \xi_d^{(t)}; \alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)})$. In iteration $t + 1$,

$$\begin{aligned} & \sum_{d=1}^D L(\gamma_d^{(t)}, \phi_d^{(t)}, \xi_d^{(t)}; \alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)}) \\ & \leq \sum_{d=1}^D L(\gamma_d^{(t+1)}, \phi_d^{(t+1)}, \xi_d^{(t+1)}; \alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)}) \\ & \leq \sum_{d=1}^D L(\gamma_d^{(t+1)}, \phi_d^{(t+1)}, \xi_d^{(t+1)}; \alpha^{(t+1)}, \Lambda^{(t+1)}, \eta_{1:c-1}^{(t+1)}). \end{aligned}$$

The first inequality holds because $(\gamma_d^{(t+1)}, \phi_d^{(t+1)}, \xi_d^{(t+1)})$ maximizes $L(\gamma_d, \phi_d, \xi_d; \alpha^{(t)}, \Lambda^{(t)}, \eta_{1:c-1}^{(t)})$ in E-step, and the second inequality holds because $(\alpha^{(t+1)}, \Lambda^{(t+1)}, \eta_{1:c-1}^{(t+1)})$ maximizes $\sum_{d=1}^D L(\gamma_d^{(t+1)}, \phi_d^{(t+1)}, \xi_d^{(t+1)}; \alpha, \Lambda, \eta_{1:c-1})$ in M step. Therefore, the objective function is guaranteed to be non-decreasing until convergence.

4.3 Prediction

Once we have the model parameters from EM, we can use $\eta_{1:c-1}$, the parameters for logistic regression, to do prediction. Given a data point $x_{1:N}$, we have

$$\begin{aligned} & E[\log p(y = h | x_{1:N}, \alpha, \Lambda, \eta_{1:c-1})] \\ & = \begin{cases} \eta_h^T E[\bar{z}] - E[\log(1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z}))] & [h]_1^{c-1} \\ 0 - E[\log(1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z}))] & h = c. \end{cases} \end{aligned}$$

Since the second term for $[h]_1^{c-1}$ and $h = c$ are the same, we only need to compare $(\eta_1^T E[\bar{z}], \dots, \eta_{c-1}^T E[\bar{z}], 0)$. If the h^{th} term is the largest, the predicted class is h .

The computation for $E[\bar{z}]$ is intractable, so we again introduce variational distribution $q(\pi, z_{1:N})$ and calculate $E_q[\bar{z}]$ as an approximation of $E[\bar{z}]$. In particular, $E_q[\bar{z}] = \frac{1}{N} \sum_{n=1}^N \phi_n$ for standard DM and $E_q[\bar{z}] = \phi$ for Fast DM.

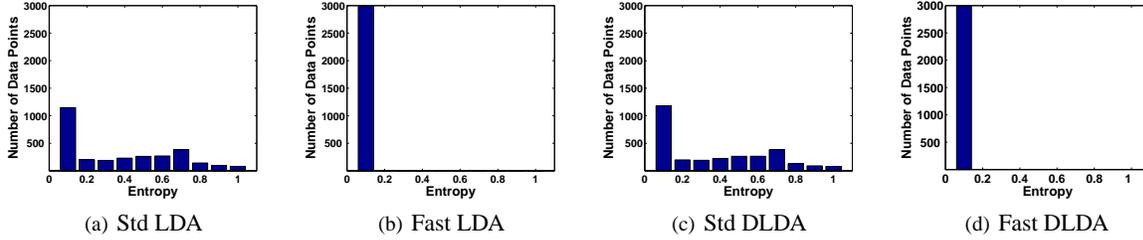


Figure 2. Histogram of Mixed-membership Entropy on Cmusim.

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Std LDA	0.9140 ±0.0140	0.6733 ±0.0254	0.9677 ±0.0069	0.8143 ± 0.0161	0.5633 ±0.0243
Fast LDA	0.9194 ±0.0148	0.6748 ± 0.0242	0.9773 ± 0.0110	0.8553 ±0.0197	0.7730 ±0.0205
Std DLDA	0.9220 ±0.0127	0.6710 ±0.0256	0.9600 ±0.0089	0.8140 ±0.0252	0.6267 ±0.0348
Fast DLDA	0.9237 ±0.0163	0.6756 ±0.0234	0.9800 ±0.0102	0.8653 ±0.0182	0.7900 ±0.0315

Table 6. Accuracy on Text with $k = c$.

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Std DLDA	549.1762 ±5.7491	2176.6794 ±(21.6241)	1752.7828 ±22.3689	2344.6408 ±966.5029	1981.4625 ±0.0348
Fast DLDA	3.6359 ±0.2090	114.3461 ±18.1366	27.569 ±0.61151	36.1029 ±2.9873	40.1892 ±5.8339

Table 7. Running time (seconds) of standard DLDA and Fast DLDA on text data with $k = c$.

5 Experimental results

We present experimental results for DMNB on UCI data, and for DLDA on text. Two types of experiments are included: First, we compare DM to corresponding MM models. Second, we compare DM with other classification algorithms. Experiments are run with a 10-fold cross validation.

5.1 Datasets

We pick 9 datasets from UCI machine learning repository for DMNB. The number of data points (D), features (N) and classes (c) in each dataset are in Table 2. We pick five text datasets for DLDA. The number of documents (D), the number of words in the dictionary (V), and the classes are in Table 3. Nasa is a subset of Aviation Safety Reporting System (ASRS) online database⁴. It contains reports of flight problems originated by three sources. Others are commonly used benchmark datasets for text classification.

5.2 DM models vs. MM models

In this section, we compare DM models with $k = c$ to corresponding MM models. We initialize model parameters

⁴<http://akama.arc.nasa.gov/ASRSDBOnline/QueryWizardBegin.aspx>

using all data points and their labels in the training set, in particular, we set the number of components k to be the number of classes c ; use the mean and standard deviation (for Gaussian case only) of the data points in each class to initialize Λ ; and use D_i/D to initialize α_i , where D_i is the number of data points in class i and D is the total number of data points. For $\eta_{1:c-1}$ in DM, we run a cross validation by holding out 10% of training data as the validation set and use the parameters generating the best results on the validation set. In particular, each η_h in $\eta_{1:c-1}$ takes value of ru_h , where u_h is a unit vector with the h^{th} dimension being 1 and others being 0, and r takes values from 0 to 100 in steps of 10. In principle, MM models are not used for classification, but given the initialization we have introduced, there is a one-to-one mapping between the component and the class. Therefore, given the mixed membership on a test data point, we pick the component i with the largest probability as the predicted component, if the corresponding class of component i is the same with the class label, we consider the data point as correctly classified, otherwise it is mistakenly classified. We use the percentage of correctly classified data points, i.e., the accuracy, to compare DM and MM.

The results for DMNB and DLDA are presented in Table 4 and 6 respectively. We make two observations: (1) Fast DM/MM models have a higher accuracy than the corresponding standard DM/MM models, with a few exceptions. (2) Standard DM models are not necessarily better than standard MM models, but Fast DM models are usually better than Fast MM models. The higher accuracy of Fast DM demonstrates the effects of logistic regression in accommodating label information for DM models.

We further investigate on the mixed memberships generated by DM and MM models. In particular, we compute the Shannon entropy for the mixed membership as a Discrete distribution and compare the entropy among different algorithms. A low entropy implies almost a “sole membership”, whereas a higher entropy implies a real mixed membership. Figure 2 is an example showing the histogram of mixed membership entropy on text data of Cmusim using four variants of LDA. We can see that for Fast LDA/DLDA, almost all data points have extremely small mixed-membership entropies, while for standard LDA/DLDA, the entropies fall into different ranges.

	Iris	Pima	Vowel	Wine	Wpbc	Ecoli	Iono	Sonar	Seg
Fast DMNB (c)	0.9600 ± 0.0644	0.7197 ± 0.0602	0.6606 ± 0.0323	0.9765 ± 0.0304	0.7632 ± 0.0832	0.8152 ± 0.0862	0.8507 ± 0.0891	0.6600 ± 0.0876	0.6701 ± 0.0487
Fast DMNB ($c+5$)	0.9600 ± 0.0716	0.7039 ± 0.0542	0.6980 ± 0.0267	0.9882 <i>± 0.0248</i>	0.7737 ± 0.0954	0.8392 ± 0.0836	0.8543 ± 0.0908	0.8100 ± 0.0907	0.7632 ± 0.0412
Fast DMNB ($c+10$)	0.9667 <i>± 0.0566</i>	0.7000 ± 0.0638	0.7020 <i>± 0.0258</i>	0.9765 ± 0.0411	0.7789 <i>± 0.1024</i>	0.8485 <i>± 0.0515</i>	0.8943 <i>± 0.0786</i>	0.8200 <i>± 0.1059</i>	0.7684 <i>± 0.0418</i>
NB	0.9533 ± 0.0632	0.7578 <i>± 0.0617</i>	0.6737 ± 0.0346	0.9705 ± 0.0310	0.7000 ± 0.0158	0.8363 ± 0.0745	0.8114 ± 0.0853	0.7268 ± 0.0079	0.6850 ± 0.0625
LR	0.9333 0.0871	0.6500 ± 0.0552	0.4515 ± 0.0444	0.7471 ± 0.1469	0.8457 ± 0.0168	0.8030 ± 0.0610	0.7171 ± 0.0494	0.5350 ± 0.0709	0.8307 ± 0.0358
SVM	0.9733 <i>± 0.0466</i>	0.7671 <i>± 0.0645</i>	0.8354 <i>± 0.0469</i>	0.9529 ± 0.0372	0.7842 <i>± 0.1323</i>	0.8394 ± 0.0670	0.9171 <i>± 0.0594</i>	0.7450 ± 0.0896	0.9745 <i>± 0.0096</i>

Table 8. Accuracy on UCI with Different Choices of k .

	Iris	Pima	Vowel	Wine	Wpbc	Ecoli	Iono	Sonar	Seg
$k=c$	0.9466 ± 0.0688	0.6400 ± 0.1577	0.4121 ± 0.0446	0.9294 ± 0.1030	0.7632 ± 0.0832	0.7666 ± 0.0655	0.7057 ± 0.1159	0.5550 ± 0.0724	0.6082 ± 0.0627
$k=c+5$	0.9400 ± 0.0913	0.6368 ± 0.1101	0.6600 ± 0.0966	0.9176 ± 0.0794	0.7631 ± 0.0917	0.7636 ± 0.0889	0.8000 ± 0.0819	0.6100 ± 0.0699	0.6346 ± 0.0734
$k=c+10$	0.9400 ± 0.0857	0.6663 ± 0.0572	0.4858 ± 0.0455	0.9235 ± 0.0411	0.7157 ± 0.1039	0.8121 ± 0.1740	0.8600 ± 0.0619	0.6450 ± 0.0845	0.6043 ± 0.0931

Table 9. Accuracy on UCI from Fast MNB and logistic regression together with different choices of k .

Similar results are obtained on UCI data. The interesting observation indicates that fast variational inference actually generates “sole membership” while standard mean-field variational inference generates real “mixed membership”. The fact that Fast DM/MM generates sole membership, as well as the previous observation that Fast DM/MM are better than standard DM/MM in terms of accuracy, shows the correlation between “sole membership” and higher classification accuracy, although we are not sure about the existence of causality between them. “Mixed membership” may be useful in various real applications, but it does not seem to help in terms of classification accuracy.

We compare the running time between standard DM and Fast DM. The results for DMNB and DLDA are presented in Table 5 and 7 respectively. In Table 5, although most of datasets are small, Fast DMNB is already faster than the standard DMNB, especially on the largest dataset Seg, where Fast DMNB is about 5 times faster than standard DMNB. Fast DM’s advantage increases when it comes to the larger and higher-dimensional text data as in Table 7, where Fast DLDA is about 20 to 150 times faster than the standard DLDA, showing Fast DM models’ absolute superiority in terms of time efficiency. Combining the results with the accuracy comparison in Table 4 and Table 6, we can see that Fast DM models are generally more accurate and substantially faster than standard DM and MM models.

5.3 Fast DM vs. other algorithms

Since Fast DM models have better performance than standard DM models, in this subsection, we use Fast DM to compare with other classification algorithms. In particular, we compare Fast DMNB with support vector machine

(SVM) [8], logistic regression (LR) and naive Bayes (NB) models on UCI data; and compare Fast DLDA with SVM, NB, LR and mixture of von Mises-Fisher (vMF) model on text data. Since DM models are combination of logistic regression and mixed-membership model, we also compare the results from DM with those from MM and logistic regression in two steps sequentially.

For Fast DM models, we run the experiments with an increasing k . In particular, for Fast DMNB, we use $k = (c, c + 5, c + 10)$, and for Fast DLDA, we use $k = (c, c + 15, c + 30, c + 50, c + 100)$. For initialization of Λ , we use the mean and standard deviation (for Gaussian case only) of the training data in given classes plus some perturbation if $k > c$; for α , we set it to be $1/k$ on each dimension; and for $\eta_{1:c-1}$, we again use a cross validation as in Section 5.2. For SVM, we use linear and RBF kernel with same cross validation strategy on the penalty parameter and the kernel parameter (for RBF only) taking values from 10^{-5} to 10^5 in multiplicative steps of 10 respectively.

The results for Fast DMNB and DLDA are presented in Table 8 and 10. The top parts of the tables are the results from the generative models, and the bottom parts are the results from discriminative classification algorithms. For SVM, we report the highest accuracy of linear and RBF kernels with different parameters. We use bold for the best results among the generative models and use bold and italic for the best results among all algorithms. Three parts of information could be read from the tables: (1) Overall, on text datasets, Fast DLDA does better than all other algorithms, including SVM, on almost all datasets, which is a promising result although more rigorous experimentations may be needed to make a further investigation; on UCI datasets, Fast DMNB also achieves higher accuracy than all other

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Fast DLDA ($k=c$)	0.9237 ± 0.0163	0.6756 ± 0.0234	0.9800 ± 0.0102	0.8653 ± 0.0182	0.7900 ± 0.0315
Fast DLDA ($k=c+15$)	0.9232 ± 0.0144	0.6858 ± 0.0216	0.9747 ± 0.0121	0.8713 ± 0.0264	0.8458 ± 0.0214
Fast DLDA ($k=c+30$)	0.9301 ± 0.0128	0.6838 ± 0.0234	0.9817 ± 0.0099	0.8707 ± 0.0228	0.8468 ± 0.0190
Fast DLDA ($k=c+50$)	0.9237 ± 0.0138	0.6854 ± 0.0211	0.9823 ± 0.0083	0.8700 ± 0.0230	0.8150 ± 0.0184
Fast DLDA ($k=c+100$)	0.9261 ± 0.0102	0.6866 ± 0.0245	0.9760 ± 0.0108	0.8718 ± 0.0182	0.8347 ± 0.0187
vMF	0.9216 ± 0.0113	0.6509 ± 0.0246	0.9530 ± 0.0071	0.7447 ± 0.0214	0.7600 ± 0.0347
NB	0.9334 ± 0.0094	0.6766 ± 0.0230	0.9813 ± 0.0069	0.8613 ± 0.0216	0.8410 ± 0.0262
LR	0.9209 ± 0.0157	0.6396 ± 0.0252	0.9553 ± 0.0157	0.6750 ± 0.1330	0.4823 ± 0.1283
SVM	0.9192 ± 0.0146	0.6854 ± 0.0278	0.9563 ± 0.0105	0.8357 ± 0.0156	0.8120 ± 0.203

Table 10. Accuracy on text with different choices of k .

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
$k=c$	0.9194 ± 0.0148	0.5609 ± 0.0281	0.9513 ± 0.0268	0.8560 ± 0.0196	0.7733 ± 0.0339
$k=c+15$	0.9118 ± 0.0124	0.5611 ± 0.0284	0.9756 ± 0.0112	0.8550 ± 0.0226	0.8173 ± 0.0197
$k=c+30$	0.9080 ± 0.0143	0.5611 ± 0.0284	0.9760 ± 0.0116	0.8530 ± 0.0216	0.8183 ± 0.0168
$k=c+50$	0.9085 ± 0.0132	0.5596 ± 0.0284	0.9746 ± 0.0123	0.8546 ± 0.0248	0.8040 ± 0.0201
$k=c+100$	0.8926 ± 0.0942	0.6537 ± 0.0598	0.9423 ± 0.0896	0.7726 ± 0.1715	0.6726 ± 0.6726

Table 11. Accuracy on text from Fast LDA and logistic regression together with different choices of k .

algorithms on most of datasets except SVM, which beats Fast DMNB six out of nine times. (2) The better performance of Fast DM models compared with LR on original datasets indicates that the low-dimensional representation we generate helps the classification. (3) Interestingly, for Fast DMNB, the accuracy increases monotonically with k from c to $c+10$ on most of the datasets. For Fast DLDA on text data, an increasing of accuracy with a larger k is also observed, although the result goes up and down without a clear trend. One possible reason for the increasing accuracy is as follows: When k is too small, we are performing a drastic dimension reduction to represent each data point in a k -dimensional mixed membership representation, which may cause a huge loss of information, but the loss may decrease when k increases.

DM models do dimensionality reduction and classification in one shot via a combination of MM models and logistic regression. In principle, we may also use these two algorithms sequentially in two steps, i.e., first using MM models to get a low-dimensional representation, and then applying logistic regression on the low-dimensional representation for classification. The results with different choices

1	runway, aircraft, approach, tower, cleared, landing, airport, turn, taxi, traffic, final, controller
2	maintenance, aircraft, flight, minimum equipment list, time, check, engine, mechanical, installed, part, inspection, work
3	passenger, flight, attendant, told, captain, seat, asked, back, attendants, aircraft, lavatory, crew
4	passenger, flight, medical, attendant, emergency, aircraft doctor, landing, attendants, captain, oxygen, paramedics

Table 12. Topics from Nasa.

of k following this two-step strategy are presented in Table 9 and 11 for UCI and text data respectively. Comparing these results with Table 8 and 10, it is clear that DM models outperform the algorithm using MM and logistic regression sequentially, which means, by combining MM and logistic regression together, DM achieves supervised dimensionality reduction to obtain a better low-dimensional representation than MM, which further helps classification. Comparing these results with the accuracy of logistic regression on original data, we can see that there is no clear winner, which may depend on the quality of low-dimensional representation generated from MM.

As we have mentioned, DM models generate interpretable results. We give an example of several topic word lists on Nasa generated by Fast DLDA ($k = c + 30$) in Table 12. It is also an interesting result demonstrating the effect of allowing a larger number of components than the number of classes, that is, Fast LDA may discover topics which are not explicitly specified in class labels. The first three topics in Table 12 correspond to three classes in Nasa respectively, but topic 4, which we call ‘‘passenger medical emergency’’, could be considered as a subcategory of the ‘‘passenger’’ class, and it is not specified in the labels. Neither NB nor SVM is able to generate this type of results.

6 Conclusion

In this paper, we have proposed discriminative mixed-membership models, as a combination of unsupervised mixed-membership models and multi-label logistic regression. We proposed a fast variational inference algorithm which is substantially faster than the mean-field approximation used in LDA. An important property of DM models is that they allow the number of components k to be different from the number of classes c . Interestingly, a larger k helps to discover the components not specified in labels and increase classification accuracy. In addition, DM models are competitive with the state of the art classification algorithms in terms of the accuracy, especially on text data, and are able to generate interpretable results. Future work includes using Dirichlet process mixture models to find out the proper value for k and extending the model to accommodate kernels.

A Variational inference

In this section, we give the derivation for variational inference in Section 4. Given the lower bound function as (7), the first five terms could easily be obtained following LDA or MNB depending on which DM model is used, so we only work on the last term $E_q[\log p(y|z_{1:N}, \eta_{1:c})]$.

The class label y is from a multi-class logistic regression LR($\frac{\exp(\eta_h^T \bar{z})}{1 + \sum_{h=1}^{k-1} \exp(\eta_h^T \bar{z})}$), $[h]_1^{k-1}$, i.e., y is from a discrete distribution with $\eta_{1:c-1} \bar{z}$ the natural parameter. Therefore,

$$p(y|z_{1:N}, \eta_{1:c-1}) = \exp\left(\sum_{h=1}^{c-1} \eta_h^T \bar{z} y_h - \log\left(1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})\right)\right).$$

Accordingly,

$$\begin{aligned} & E_q[\log p(y|z_{1:N}, \eta_{1:c-1})] \\ &= E_q\left[\sum_{h=1}^{c-1} \eta_h^T \bar{z} y_h - \log\left(1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})\right)\right] \\ &= \sum_{h=1}^{c-1} \sum_{i=1}^k \eta_{hi} E_q[\bar{z}_i] y_h - E_q\left[\log\left(1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})\right)\right]. \end{aligned} \quad (12)$$

The second term of (12) could be expanded as follows:

$$\begin{aligned} & - E_q\left[\log\left(1 + \sum_{h=1}^{c-1} \exp(\eta_h^T \bar{z})\right)\right] \\ & \geq - \log\left(1 + \sum_{h=1}^{c-1} E_q\left[\exp\left(\sum_{i=1}^k \eta_{hi} \bar{z}_i\right)\right]\right) \\ & \geq - \log\left(1 + \sum_{h=1}^{c-1} E_q\left[\sum_{i=1}^k \bar{z}_i \exp(\eta_{hi})\right]\right) \\ & = - \log\left(1 + \sum_{h=1}^{c-1} \sum_{i=1}^k E_q[\bar{z}_i] \exp(\eta_{hi})\right) \\ & \geq - \frac{1}{\xi} \sum_{h=1}^{c-1} \sum_{i=1}^k E_q[\bar{z}_i] \exp(\eta_{hi}) + 1 - \frac{1}{\xi} - \log(\xi), \end{aligned} \quad (13)$$

where the first inequality is from Jensen's inequality, the second inequality is also from Jensen's inequality noticing that \bar{z} is actually a Discrete distribution, and the third inequality is from $-\log(x) \geq 1 - \frac{x}{\xi} - \log(\xi)$ [16] by introducing a new variational parameter $\xi > 0$. Given (13),

$$\begin{aligned} & E_q[\log p(y|z_{1:N}, \eta_{1:c-1})] \\ & \geq \sum_{i=1}^k E_q[\bar{z}_i] \sum_{h=1}^{c-1} \left(\eta_{hi} y_h - \frac{1}{\xi} \exp(\eta_{hi})\right) + 1 - \frac{1}{\xi} - \log(\xi), \end{aligned}$$

where in standard DM models $E_q[\bar{z}_i] = \frac{1}{N} \sum_{n=1}^N \phi_{ni}$, and in fast DM models, $E_q[\bar{z}_i] = \phi_i$.

Putting $E_q[\log p(y|z_{1:N}, \eta_{1:c-1})]$ back to (7) gives us the complete expression for L . By maximizing (7) with respect to the variational and model parameters alternatively as in Section 4, we find the optimal value for $(\alpha, \Lambda, \eta_{1:c-1})$.

Acknowledgements

The research was supported by NASA grant NNX08AC36A and NSF grant IIS-0812183.

References

- [1] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *JMLR*, 9:1823–1856, 2008.
- [2] A. Banerjee and H. Shan. Latent Dirichlet conditional naive Bayes models. In *ICDM*, 2007.
- [3] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley and Sons, 1978.
- [4] D. Blei and M. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.
- [5] D. Blei and J. McAuliffe. Supervised topic models. In *NIPS*, 2007.
- [6] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.
- [7] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [8] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [9] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning Journal*, 29:103–130, 1997.
- [10] P. Flaherty, G. Giaever, M. Jordan, and A. Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21:3286–3293, 2005.
- [11] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5225, 2004.
- [12] P. Koutsourelakis and T. Eliassi-Rad. Finding mixed-memberships in social networks. In *AAAI*, 2008.
- [13] S. Lacoste-Julien, F. Sha, and M. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.
- [14] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.
- [15] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, 2008.
- [16] T. Minka. A comparison of numerical optimizers for logistic regression. Technical report, 2003.
- [17] F. Pampel. *Logistic Regression: A Primer*. Sage, 2000.
- [18] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.
- [19] H. Wang, M. Huang, and X. Zhu. A generative probabilistic model for multi-label classification. In *ICDM*, 2008.