# Using Both Latent and Supervised Shared Topics for Multitask Learning

Ayan Acharya[1], Aditya Rawal[2], Raymond J. Mooney[2], and Eduardo R. Hruschka[3]

[1] Department of ECE, University of Texas at Austin, USA
aacharya@utexas.edu
[2] Department of CS, University of Texas at Austin, USA
{aditya,mooney}@cs.utexas.edu
[3] Department of CS, University of São Paulo at São Carlos, Brazil
erh@icmc.usp.br

**Abstract.** This paper introduces two new frameworks, Doubly Supervised Latent Dirichlet Allocation (DSLDA) and its non-parametric variation (NP-DSLDA), that integrate two different types of supervision: topic labels and category labels. This approach is particularly useful for multitask learning, in which both latent and supervised topics are shared between multiple categories. Experimental results on both document and image classification show that both types of supervision improve the performance of both DSLDA and NP-DSLDA and that sharing both latent *and* supervised topics allows for better multitask learning.

## 1 Introduction

Humans can distinguish as many as 30,000 relevant object classes [7]. Training an isolated object detector for each of these different classes would require millions of training examples in aggregate. Computer vision researchers have proposed a more efficient learning mechanism in which object categories are learned via *shared* attributes, abstract descriptors of object properties such as "striped" or "has four legs" [17,25,24]. The attributes serve as an intermediate layer in a classifier cascade. The classifier in the first stage is trained to predict the attributes from the raw features and that in the second stage is trained to predict the categories from the attributes. During testing, only the raw features are observed and the attributes must be inferred. This approach is inspired by human perception and learning from high-level object descriptions. For example, from the phrase "eight-sided red traffic sign with white writing", humans can detect stop signs [25]. Similarly, from the description "large gray animals with long trunks", human can identify elephants. If the *shared* attributes transcend object class boundaries, such a classifier cascade is beneficial for *transfer learning* [28] where fewer labeled examples are available for some object categories compared to others [25].

Multitask learning (MTL) is a form of transfer learning in which simultaneously learning multiple related "tasks" allows each one to benefit from the learning of all of the others. If the tasks are related, training one task should provide helpful "inductive bias" for learning the other tasks. To enable the reuse of training information across multiple related tasks, all tasks might utilize the same latent shared intermediate representation – for example, a shared hidden layer in a multi-layer perceptron [11]. In this

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part II, LNAI 8189, pp. 369–384, 2013.

case, the training examples for all tasks provide good estimates of the weights connecting the input layer to the hidden layer, and hence only a small number of examples per task is sufficient to achieve high accuracy. This approach is in contrast to "isolated" training of tasks where each task is learned independently using a separate classifier.

In this paper, our objective is to combine these two approaches to build an MTL framework that can use *both* attributes *and* class labels. The multiple tasks here correspond to different object categories (classes), and *both* observable attributes and latent properties are shared across the tasks. We want to emphasize that the proposed frameworks support general MTL; however, the datasets we use happen to be multiclass, where each class is treated as a separate "task" (as typical in multi-class learning based on binary classifiers). But, in no way are the frameworks restricted to multiclass MTL. Since attribute-based learning has been shown to support effective transfer learning in computer vision, the tasks here naturally correspond to object classes.

The basic building block of the frameworks presented in this paper is Latent Dirichlet Allocation (LDA) [9]. LDA focuses on unsupervised induction of multiple "topics" that help characterize a corpus of text documents. LDA has also been applied in computer vision where SIFT features are appropriately quantized to generate a *bag of visual words* for representing an image [35]. Since our experiments use both text and image data, we will overload the word "document" to denote either a text document or an image represented as a bag of visual words. The LDA approach has also been augmented to include two different types of supervision, document-level labels for either topics [31] or for an overall category inferred from the topics [43]. This paper introduces two new approaches, Doubly Supervised Latent Dirichlet Allocation (DSLDA) and its non-parametric variation (NP-DSLDA), that integrate both forms of supervision. At the topic level, the models assume that supervision is available for some topics during training (corresponding to the "attributes" used in computer vision), but that other topics remain latent (corresponding to the hidden layer in traditional MTL). The ability to provide supervision for *both* categories and a *subset* of topics improves the models' ability to perform accurate classification. In many applications, a variety of kinds of supervision may be naturally available from different sources at multiple levels of abstraction, such as keywords, topics, and categories for documents, or visual attribute, object, and scene labels for images. By effectively utilizing such multiple, interacting levels of supervision, DSLDA is able to learn more accurate predictors. In a supervised LDA [8,43] setting, forcing multiple tasks to share the same set of latent topics results in an LDA-based approach to MTL. By allowing supervision to also be provided for a subset of these shared topics, DSLDA and NP-DSLDA support a particularly effective form of MTL.

The rest of the paper is organized as follows. We present related literature in Section 2, followed by the descriptions of DSLDA and NP-DSLDA in Section 3 and Section 4 respectively. Experimental results on both multi-class image and document categorization are presented in Section 5, demonstrating the value of integrating both supervised and latent shared topics in diverse applications. Finally, future directions and conclusions are presented in Section 6.

**Note on Notation:** Vectors and matrices are denoted by bold-faced lowercase and capital letters, respectively. Scalar variables are written in italic font, and sets are denoted

by calligraphic uppercase letters. Dir(), Beta() and multinomial() stand for Dirichlet, Beta and multinomial distribution respectively.

## 2 Related Work

### 2.1 Statistical Topic Models

LDA [9] treats documents as a mixture of topics, which in turn are defined by a distribution over a set of words. The words in a document are assumed to be sampled from multiple topics. In its original formulation, LDA can be viewed as a purely-unsupervised form of dimensionality reduction and clustering of documents in the topic space, although several extensions of LDA have subsequently incorporated some sort of supervision. Some approaches provide supervision by labeling each document with its set of topics [31,32]. In particular, in *Labeled LDA* (LLDA [31]), the primary objective is to build a model of the words that indicate the presence of certain topic labels. For example, when a user explores a webpage based on certain tags, LLDA can be used to highlight interesting portions of the page or build a summary of the text from multiple webpages that share the same set of tags. The words in a given training document are assumed to be sampled *only* from the supervised topics, which the document has been labeled as covering.

Some other researchers [8,43,12] assume that supervision is provided for a single *response variable* to be predicted for a given document. The response variable might be real-valued or categorical, and modeled by a normal, Poisson, Bernoulli, multinomial or other distribution (see [12] for details). Some examples of documents with response variables are essays with their grades, movie reviews with their numerical ratings, web pages with their number of hits over a certain period of time, and documents with category labels. In *Maximum Entropy Discriminative LDA* (MedLDA) [43], the objective is to infer some low-dimensional (topic-based) representation of documents which is predictive of the response variable. Essentially, MedLDA solves two problems jointly – dimensionality reduction and max-margin classification using the features in the dimensionally-reduced space. Compared to earlier versions of supervised topic models [8,12], MedLDA has simpler update equations and produces superior experimental results. Therefore, in the frameworks presented in Sections 3.2 and 4, the max-margin principle adopted in MedLDA is preferred over other supervised topic models.

### 2.2 Transfer and Multitask Learning

Transfer learning allows the learning of some tasks to benefit the learning of others through either simultaneous [11] or sequential [10] training. In multitask learning (MTL [11]), a single model is simultaneously trained to perform multiple related tasks. MTL has emerged as a very promising research direction for various applications including biomedical informatics [6], marketing [15], natural language processing [2], and computer vision [34].

Many different MTL approaches have been proposed over the past 15 years (*e.g.*, see [38,28,29] and references therein). These include different learning methods, such

as empirical risk minimization using group-sparse regularizers [20,23,21], hierarchical Bayesian models [41,26] and hidden conditional random fields [30]. Evgeniou *et al.* [14] proposed the regularized MTL which constrained the models of all tasks to be close to each other. The task relatedness in MTL has also been modeled by constraining multiple tasks to share a common underlying structure [5,3,11]. Ando and Zhang [1] proposed a structural learning formulation, which assumed multiple predictors for different tasks shared a common structure on the underlying predictor space.

In all of the MTL formulations mentioned above, the basic assumption is that all tasks are related. In practical applications, these might not be the case and the tasks might exhibit a more sophisticated group structure. Such structure is handled using clustered multi-task learning (CMTL). In [4] CMTL is implemented by considering a mixture of Gaussians instead of single Gaussian priors. Xue *et al.* [39] introduced the Dirichlet process prior that automatically identifies subgroups of related tasks. In [19], a clustered MTL framework was proposed that simultaneously identified clusters and performed multi-task inference.

In the models presented in the next two sections, an LDA-based approach to MTL is easily obtained by maintaining a common set of topics to support the prediction of multiple response variables. This idea is analogous to implementing MTL using a common shared underlying structure [5,3,11]. We will also explain how NP-DSLDA is capable of performing CMTL.

## 3 Doubly Supervised LDA (DSLDA)

### 3.1 Task Definition

Assume we are given a training corpus consisting of $N$ documents belonging to $Y$ different classes (where each document belongs to exactly one class and each class corresponds to a different task). Further assume that each of these training documents is also annotated with a set of $K_2$ different topic "tags" (henceforth referred to as "supervised topics"). For computer vision data, the supervised topics correspond to the attributes provided by human experts. The objective is to train a model using the words in a data, as well as the associated supervised topic tags and class labels, and then use this model to classify completely unlabeled test data for which no topic tags nor class labels are provided. The human-provided supervised topics are presumed to provide abstract information that is helpful in predicting the class labels of test documents.

### 3.2 Generative Model

In order to include both types of supervision (class and topic labels), a combination of the approaches described in Section 2.1 is proposed. Note that LLDA uses *only* supervised topics and does not have any mechanism for generating class labels. On the other hand, MedLDA has only *latent* topics but learns a discriminative model for predicting classes from these topics. To the best of our knowledge, ours is the first LDA approach to integrate both types of supervision in a single framework. The generative process of DSLDA is described below.

For the $n^{\text{th}}$ document, sample a topic selection probability vector $\boldsymbol{\theta}_n \sim \text{Dir}(\boldsymbol{\alpha}_n)$, where $\boldsymbol{\alpha}_n = \boldsymbol{\Lambda}_n \boldsymbol{\alpha}$ and $\boldsymbol{\alpha}$ is the parameter of a Dirichlet distribution of dimension $K$, which is the total number of topics. The topics are assumed to be of two types – latent and supervised, and there are $K_1$ latent topics and $K_2$ supervised topics. Therefore, $K = K_1 + K_2$. Latent topics are never observed, while supervised topics are observed in training but not in test data. Henceforth, in each vector or matrix with $K$ components, it is assumed that the first $K_1$ components correspond to the latent topics and the next $K_2$ components to the supervised topics. $\boldsymbol{\Lambda}_n$ is a diagonal binary matrix of dimension $K \times K$. The $k^{\text{th}}$ diagonal entry is unity if *either* $1 \leq k \leq K_1$ *or* $K_1 < k \leq K$ and the $n^{\text{th}}$ document is tagged with the $k^{\text{th}}$ topic. Also, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ where $\boldsymbol{\alpha}_1$ is a parameter of a Dirichlet distribution of dimension $K_1$ and $\boldsymbol{\alpha}_2$ is a parameter of a Dirichlet distribution of dimension $K_2$.

For the $m^{\text{th}}$ word in the $n^{\text{th}}$ document, sample a topic $z_{nm} \sim \text{multinomial}(\boldsymbol{\theta}'_n)$, where $\boldsymbol{\theta}'_n = (1-\epsilon)\{\boldsymbol{\theta}_{nk}\}_{k=1}^{k_1} \epsilon \{\Lambda_{n,kk}\boldsymbol{\theta}_{nk}\}_{k=1+k_1}^{K}$. This implies that the supervised topics are weighted by $\epsilon$ and the latent topics are weighted by $(1-\epsilon)$. Sample the word $w_{nm} \sim \text{multinomial}(\boldsymbol{\beta}_{z_{nm}})$, where $\boldsymbol{\beta}_k$ is a multinomial distribution over the vocabulary of words corresponding to the $k^{\text{th}}$ topic.

For the $n^{\text{th}}$ document, generate $Y_n = \arg\max_y \boldsymbol{r}_y^T \mathbb{E}(\bar{\mathbf{z}}_n)$ where $Y_n$ is the class label associated with the $n^{\text{th}}$ document, $\bar{\mathbf{z}}_n = \sum_{m=1}^{M_n} \mathbf{z}_{nm}/M_n$. Here, $\mathbf{z}_{nm}$ is an indicator vector of dimension $K$. $\boldsymbol{r}_y$ is a $K$-dimensional real vector corresponding to the $y^{\text{th}}$ class, and it is assumed to have a prior distribution $\mathcal{N}(0, 1/C)$. $M_n$ is the number of words in the $n^{\text{th}}$ document. The maximization problem to generate $Y_n$ (or the classification problem) is carried out using a max-margin principle.

Note that predicting each class is effectively treated as a separate task, and that the shared topics are useful for generalizing the performance of the model across classes. In particular, when all classes have few training examples, knowledge transfer between classes can occur through the shared topics. So, the mapping from the original feature space to the topic space is effectively learned using examples from all classes, and a few examples from each class are sufficient to learn the mapping from the reduced topic space to the class labels.

### 3.3 Inference and Learning

Let us denote the hidden variables by $\boldsymbol{Z} = \{\{z_{nm}\}, \{\boldsymbol{\theta}_n\}\}$, the observed variables by $\boldsymbol{X} = \{w_{nm}\}$ and the model parameters by $\boldsymbol{\kappa}_0$. The joint distribution of the hidden and observed variables is:

$$p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\kappa}_0) = \prod_{n=1}^{N} p(\boldsymbol{\theta}_n|\boldsymbol{\alpha}_n) \prod_{m=1}^{M_n} p(z_{nm}|\boldsymbol{\theta}'_n) p(w_{nm}|\boldsymbol{\beta}_{\boldsymbol{z}_{nm}}) \tag{1}$$

To avoid computational intractability, inference and estimation are performed using Variational **EM**. The factorized approximation to the posterior distribution on hidden variables $\mathbf{Z}$ is given by:

$$q(\boldsymbol{Z}|\{\boldsymbol{\kappa}_n\}_{n=1}^{N}) = \prod_{n=1}^{N} q(\boldsymbol{\theta}_n|\boldsymbol{\gamma}_n) \prod_{m=1}^{M_n} q(z_{nm}|\boldsymbol{\phi}_{nm}), \tag{2}$$

where $\boldsymbol{\theta}_n \sim \text{Dir}(\boldsymbol{\gamma}_n)\ \forall n \in \{1, 2, \cdots, N\}$, $z_{nm} \sim \text{multinomial}(\boldsymbol{\phi}_{nm})\ \forall n \in \{1, 2, \cdots, N\}$ and $\forall m \in \{1, 2, \cdots, M_n\}$, and $\boldsymbol{\kappa}_n = \{\boldsymbol{\gamma}_n, \{\boldsymbol{\phi}_{nm}\}\}$, which is the set of variational parameters corresponding to the $n^{\text{th}}$ instance. Further, $\boldsymbol{\gamma}_n = (\gamma_{nk})_{k=1}^K\ \forall n$, and $\boldsymbol{\phi}_{nm} = (\phi_{nmk})_{k=1}^K\ \forall n, m$. With the use of the lower bound obtained by the factorized approximation, followed by Jensen's inequality, DSLDA reduces to solving the following optimization problem[1]:

$$\min_{q, \boldsymbol{\kappa}_0, \{\xi_n\}} \frac{1}{2}||\boldsymbol{r}||^2 - \mathcal{L}(q(\boldsymbol{Z})) + C\sum_{n=1}^{N}\xi_n,$$
$$\text{s.t. } \forall n, y \neq Y_n : \mathbb{E}[\boldsymbol{r}^T \Delta f_n(y)] \geq 1 - \xi_n; \xi_n \geq 0. \tag{3}$$

Here, $\Delta f_n(y) = f(Y_n, \bar{\boldsymbol{z}}_n) - f(y, \bar{\boldsymbol{z}}_n)$ and $\{\xi_n\}_{n=1}^N$ are the slack variables, and $f(y, \bar{\boldsymbol{z}}_n)$ is a feature vector whose components from $(y-1)K+1$ to $yK$ are those of the vector $\bar{\boldsymbol{z}}_n$ and all the others are 0. $\mathbb{E}[\boldsymbol{r}^T \Delta f_n(y)]$ is the "expected margin" over which the true label $Y_n$ is preferred over a prediction $y$. From this viewpoint, DSLDA projects the documents onto a combined topic space and then uses a max-margin approach to predict the class label. The parameter $C$ penalizes the margin violation of the training data.

$$\phi^*_{nmk} \propto \Lambda_{n,kk}\exp\left[\psi(\gamma_{nk}) + \log(\beta_{kw_{nm}}) + \log(\epsilon')\right. \tag{4}$$
$$\left. +1/M_n \sum_{y \neq Y_n} \mu_n(y)\mathbb{E}[r_{Y_n k} - r_{yk}]\right]\ \forall n, m, k.$$

$$\gamma^*_{nk} = \Lambda_{n,kk}\left[\alpha_k + \sum_{m=1}^{M_n}\phi_{nmk}\right]\ \forall n, vk. \tag{5}$$

$$\beta^*_{kv} \propto \sum_{n=1}^{N}\sum_{m=1}^{M_n}\phi_{nmk}\mathbb{I}_{\{w_{nm}=v\}}\ \forall k, v. \tag{6}$$

$$\mathcal{L}_{[\boldsymbol{\alpha}_1/\boldsymbol{\alpha}_2]} = \left[\sum_{n=1}^{N}\log(\Gamma(\sum_{k=1}^{K}\alpha_{nk})) - \sum_{n=1}^{N}\sum_{k=1}^{K}\log(\Gamma(\alpha_{nk}))\right] \tag{7}$$
$$+\sum_{n=1}^{N}\sum_{k=1}^{K}\left[\psi(\gamma_{nk}) - \psi(\sum_{k=1}^{K}\gamma_{nk})\right](\alpha_{nk} - 1).$$

Let $\mathcal{Q}$ be the set of all distributions having a fully factorized form as given in (2). Let the distribution $q^*$ from the set $\mathcal{Q}$ optimize the objective in Eq. (3). The optimal values of corresponding variational parameters are given in Eqs. (4) and (5). In Eq. (4), $\epsilon' = (1-\epsilon)$ if $k \leq K_1$ and $\epsilon' = \epsilon$ otherwise. Since $\boldsymbol{\phi}_{nm}$ is a multinomial distribution, the updated values of the $K$ components should be normalized to unity. The optimal values of $\boldsymbol{\phi}_{nm}$ depend on $\boldsymbol{\gamma_n}$ and vice-versa. Therefore, iterative optimization is adopted to maximize the lower bound until convergence is achieved.

[1] Please see [43] for further details.

During testing, one does not observe a document's supervised topics and, in principle, has to explore $2^{K_2}$ possible combinations of supervised tags – an expensive process. A simple approximate solution, as employed in LLDA [31], is to assume the absence of the variables $\{\boldsymbol{\Lambda}_n\}$ altogether in the test phase, and just treat the problem as inference in MedLDA with $K$ latent topics. One can then threshold over the last $K_2$ topics if the tags of a test document need to be inferred. Equivalently, one can also assume $\boldsymbol{\Lambda}_n$ to be an identity matrix of dimension $K \times K \; \forall n$. This representation ensures that the expressions for update equations (4) and (5) do not change in the test phase.

In the M step, the objective in Eq. (3) is maximized w.r.t $\boldsymbol{\kappa}_0$. The optimal value of $\beta_{kv}$ is given in Eq. (6). Since $\boldsymbol{\beta}_k$ is a multinomial distribution, the updated values of the $V$ components should be normalized. However, numerical methods for optimization are required to update $\boldsymbol{\alpha}_1$ or $\boldsymbol{\alpha}_2$. The part of the objective function that depends on $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ is given in Eq. (7). The update for the parameter $\boldsymbol{r}$ is carried out using a multi-class SVM solver [16]. With all other model and variational parameters held fixed (*i.e.* with $\mathcal{L}(q)$) held constant), the objective in Eq. (3) is optimized w.r.t. $\boldsymbol{r}$. A reader familiar with the updates in unsupervised LDA can see the subtle (but non-trivial) changes in the update equations for DSLDA.

## 4 Non-parametric DSLDA

We now propose a non-parametric extension of DSLDA (NP-DSLDA) that solves the model selection problem and automatically determines the best number of latent topics for modeling the given data. A modified stick breaking construction of Hierarchical Dirichlet Process (HDP) [33], recently introduced in [36] is used here which makes variational inference feasible. The idea in such representation is to share the corpus level atoms across documents by sampling atoms with replacement for each document and modifying the weights of these samples according to some other GEM distribution [33] whose parameter does not depend on the weights of the corpus-level atoms.

The combination of an infinite number of latent topics with a finite number of supervised topics in a single framework is not trivial and ours is the first model to accomplish this. One simpler solution is to introduce one extra binary hidden variable for each word in each document which could select either the set of latent topics or the set of supervised topics. Subsequently, a word in a document can be sampled from either the supervised or the latent topics based on the value sampled by the hidden "switching" variable. However, the introduction of such extra hidden variables adversely affects model performance as explained in [13]. In NP-DSLDA, we are able to avoid such extra hidden variables by careful modeling of the HDP. This will be evident in the generative process of NP-DSLDA presented below:

- Sample $\boldsymbol{\phi}_{k_1} \sim \text{Dir}(\boldsymbol{\eta}_1) \forall k_1 \in \{1, 2, \cdots, \infty\}$ and $\boldsymbol{\phi}_{k_2} \sim \text{Dir}(\boldsymbol{\eta}_2) \; \forall k_2 \in \{1, 2, \cdots, K_2\}$. $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ are the parameters of Dirichlet distribution of dimension $V$.
- Sample $\beta'_{k_1} \sim \text{Beta}(1, \delta_0) \; \forall k_1 \in \{1, 2, \cdots, \infty\}$.
- For the $n^{\text{th}}$ document, sample $\boldsymbol{\pi}_n^{(2)} \sim \text{Dir}(\boldsymbol{\Lambda}_n \boldsymbol{\alpha}_2)$. $\boldsymbol{\alpha}_2$ is the parameter of Dirichlet of dimension $K_2$. $\boldsymbol{\Lambda}_n$ is a diagonal binary matrix of dimension $K_2 \times K_2$. The $k^{\text{th}}$ diagonal entry is unity if the $n^{\text{th}}$ word is tagged with the $k^{\text{th}}$ supervised topic.

- $\forall n, \forall t \in \{1, 2, \cdots, \infty\}$, sample $\pi'_{nt} \sim \text{Beta}(1, \alpha_0)$. Assume $\boldsymbol{\pi}_n^{(1)} = (\pi_{nt})_t$ where $\pi_{nt} = \pi'_{nt} \prod_{l<t}(1 - \pi'_{nl})$.
- $\forall n, \forall t$, sample $c_{nt} \sim \text{multinomial}(\boldsymbol{\beta})$ where $\beta_{k_1} = \beta'_{k_1} \prod_{l<k_1}(1 - \beta'_l)$. $\boldsymbol{\pi}_n^{(1)}$ represents the probability of selecting the sampled atoms in $\boldsymbol{c}_n$. Due to sampling with replacement, $\boldsymbol{c}_n$ can contain multiple atoms of the same index from the corpus level DP.
- For the $m^{\text{th}}$ word in the $n^{\text{th}}$ document, sample $z_{nm} \sim \text{multinomial}((1 - \epsilon)\boldsymbol{\pi}_n^{(1)}, \epsilon\boldsymbol{\pi}_n^{(2)})$. This implies that w.p. $\epsilon$, a topic is selected from the set of supervised topics and w.p. $(1 - \epsilon)$, a topic is chosen from the set of (infinite number of) unsupervised topics. Note that by weighting the $\boldsymbol{\pi}$'s appropriately, the need for additional hidden "switching" variable is avoided.
- Sample $w_{nm}$ from a multinomial given by the following equation:

$$\prod_{k_1=1}^{\infty} \prod_{v=1}^{V} \phi_{k_1 v}^{\mathbb{I}_{\{w_{nm}=v\}}\mathbb{I}_{\{c_{nz_{nm}}=k_1 \in \{1,\cdots,\infty\}\}}} \prod_{k_2=1}^{K_2} \prod_{v=1}^{V} \phi_{k_2 v}^{\mathbb{I}_{\{w_{nm}=v\}}\mathbb{I}_{\{z_{nm}=k_2 \in \{1,\cdots,K_2\}\}}} \quad (8)$$

The joint distribution of NP-DSLDA is given as follows:

$$p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\kappa}_0) = \prod_{k_1=1}^{\infty} p(\phi_{k_1}|\boldsymbol{\eta}_1) p(\beta'_{k_1}|\boldsymbol{\delta}_0) \prod_{k_2=1}^{K_2} p(\phi_{k_2}|\boldsymbol{\eta}_2) \prod_{n=1}^{N} p(\boldsymbol{\pi}_n^{(2)}|\boldsymbol{\alpha}_2) \quad (9)$$

$$\prod_{t=1}^{\infty} p(\boldsymbol{\pi}_{nt}^{\prime(1)}|\boldsymbol{\alpha}_0) p(c_{nt}|\boldsymbol{\beta}') \prod_{m=1}^{M_n} p(z_{nm}|\boldsymbol{\pi}_n^{(1)}, \boldsymbol{\pi}_n^{(2)}, \epsilon) p(w_{nm}|\boldsymbol{\phi}, c_{nz_{nm}}, z_{nm}).$$

As an approximation to the posterior distribution over the hidden variables, we use the following factorized distribution:

$$q(\boldsymbol{Z}|\boldsymbol{\kappa}) = \prod_{k_1=1}^{\overline{K_1}} q(\phi_{k_1}|\boldsymbol{\lambda}_{k_1}) \prod_{k_2=1}^{K_2} q(\phi_{k_2}|\boldsymbol{\lambda}_{k_2}) \prod_{k_1=1}^{\overline{K_1}-1} q(\beta'_{k_1}|u_{k_1}, v_{k_1}) \quad (10)$$

$$\prod_{n=1}^{N} q(\boldsymbol{\pi}_n^{(2)}|\boldsymbol{\gamma}_n) \prod_{t=1}^{T-1} q(\boldsymbol{\pi}_{nt}^{\prime(1)}|a_{nt}, b_{nt}) \prod_{t=1}^{T} q(c_{nt}|\boldsymbol{\varphi}_{nt}) \prod_{m=1}^{M_n} q(z_{nm}|\boldsymbol{\zeta}_{nm}).$$

Here, $\boldsymbol{\kappa}_0$ and $\boldsymbol{\kappa}$ denote the sets of model and variational parameters, respectively. $\overline{K_1}$ is the truncation limit of the corpus-level Dirichlet Process and $T$ is the truncation limit of the document-level Dirichlet Process. $\{\boldsymbol{\lambda}_k\}$ are the parameters of Dirichlet each of dimension $V$. $\{u_{k_1}, v_{k_1}\}$ and $\{a_{nt}, b_{nt}\}$ are the parameters of variational Beta distribution corresponding to corpus level and document level sticks respectively. $\{\boldsymbol{\varphi}_{nt}\}$ are multinomial parameters of dimension $\overline{K_1}$ and $\{\boldsymbol{\zeta}_{nm}\}$ are multinomials of dimension $(T + K_2)$. $\{\boldsymbol{\gamma}_n\}_n$ are parameters of Dirichlet distribution of dimension $K_2$.

The underlying optimization problem takes the same form as in Eq. (3). The only difference lies in the calculation of $\Delta f_n(y) = f(Y_n, \bar{\boldsymbol{s}}_n) - f(y, \bar{\boldsymbol{s}}_n)$. The first set of dimensions of $\bar{\boldsymbol{s}}_n$ (corresponding to the unsupervised topics) is given by $1/M_n \sum_{m=1}^{M_n} \boldsymbol{c}_{nz_{nm}}$, where $\boldsymbol{c}_{nt}$ is an indicator vector over the set of unsupervised topics. The following $K_2$ dimensions (corresponding to the supervised topics) are given by $1/M_n \sum_{m=1}^{M_n} \boldsymbol{z}_{nm}$. After the variational approximation with $\overline{K_1}$ number of corpus level sticks, $\bar{\boldsymbol{s}}_n$ turns out

to be of dimension $(\overline{K_1}+K_2)$ and the feature vector $f(y,\bar{s}_n)$ constitutes $Y(\overline{K_1}+K_2)$ elements. The components of $f(y,\bar{s}_n)$ from $(y-1)(\overline{K_1}+K_2)+1$ to $y(\overline{K_1}+K_2)$ are those of the vector $\bar{s}_n$ and all the others are 0. Essentially, due to the variational approximation, NP-DSLDA projects each document on to a combined topic space of dimension $(\overline{K_1}+K_2)$ and learns the mapping from this space to the classes.

$$\zeta^*_{nmt} \propto \exp\Bigg[[\psi(a_{nt})-\psi(a_{nt}+b_{nt})]\mathbb{I}_{\{t<T\}}+\sum_{t'=1}^{t-1}[\psi(b_{nt'})-\psi(a_{nt'}+b_{nt'})] + \sum_{k_1=1}^{\overline{K_1}}\varphi_{ntk_1}\left[\psi(\lambda_{k_1 w_{nm}})-\psi(\sum_{v=1}^{V}\lambda_{k_1 v})\right] + \sum_{y\neq Y_n}\mu_n(y)\sum_{k_1=1}^{\overline{K_1}}\mathbb{E}[r_{Y_n k_1}-r_{yk_1}]\varphi_{ntk_1}\Bigg] \quad \forall n,m,t. \tag{11}$$

$$\zeta^*_{nm(T+k_2)} \propto \Lambda_{nk_2k_2}\exp\Bigg[\psi(\gamma_{nk_2})-\psi(\sum_{k_2=1}^{K_2}\gamma_{nk_2})+\psi(\lambda_{(\overline{K_1}+k_2)w_{nm}}) -\psi(\sum_{v=1}^{V}\lambda_{(\overline{K_1}+k_2)v})+1/M_n\sum_{y\neq Y_n}\mu_n(y)\mathbb{E}[r_{Y_n(\overline{K_1}+k_2)}-r_{y(\overline{K_1}+k_2)}]\Bigg] \quad \forall n,m,k_2. \tag{12}$$

$$\varphi^*_{ntk_1} \propto \exp\left[[\psi(u_{k_1})-\psi(u_{k_1}+v_{k_1})]\,\mathbb{I}_{\{k_1<K_1\}}\right. + \sum_{k'=1}^{k_1-1}[\psi(v_{k'})-\psi(u_{k'}+v_{k'})]+\sum_{m=1}^{M_n}\zeta_{nmt}\left[\psi(\lambda_{k_1 w_{nm}})-\psi(\sum_{v=1}^{V}\lambda_{k_1 v})\right] \left.+1/M_n\sum_{y\neq Y_n}\mu_n(y)\mathbb{E}[r_{Y_n k_1}-r_{yk_1}]\left(\sum_{m=1}^{M_n}\zeta_{nmt}\right)\right] \quad \forall n,t,k_1. \tag{13}$$

Some of the update equations of NP-DSLDA are given in the above equations, where $\{\varphi_{ntk_1}\}$ are the set of variational parameters that characterize the assignment of the documents to the global set of $(\overline{K_1}+K_2)$ topics. One can see how the effect of the class labels is included in the update equation of $\{\varphi_{ntk_1}\}$ via the average value of the parameters $\{\zeta_{nmt}\}$. This follows intuitively from the generative assumption. update exists for the model parameters and hence numerical optimization has to be used. Other updates are either similar to DSLDA or the model in [36] and are omitted due to space constraints. $\{\boldsymbol{\zeta}_{nm}\}$, corresponding to supervised and unsupervised topics, should be individually normalized and then scaled by $\epsilon$ and $(1-\epsilon)$ respectively. Otherwise, the effect of the Dirichlet prior on supervised topics will get compared to that of the GEM prior on the unsupervised topics which does not follow the generative assumptions. The variational parameters $\{\boldsymbol{\lambda}_k\}$ and $\{\boldsymbol{\varphi}_{nt}\}$ are also normalized.

Note that NP-DSLDA offers some flexibility with respect to the latent topics that could be dominant for a specific task. One could therefore postulate that NP-DSLDA can learn the clustering of tasks from the data itself by making a subset of latent topics to be dominant for a set of tasks. Although do not have supporting experiments, NP-DSLDA is capable in principle of performing clustered multi-task learning without any prior assumption on the relatedness of the tasks.

## 5 Experimental Evaluation

### 5.1 Data Description

Our evaluation used two datasets, a text corpus and a multi-class image database, as described below.

**aYahoo Data.** The first set of experiments was conducted with the aYahoo image dataset from [17] which has 12 classes – carriage, centaur, bag, building, donkey, goat, jetski, monkey, mug, statue, wolf, and zebra.[2] Each image is annotated with relevant visual attributes such as "has head", "has wheel", "has torso" and 61 others, which we use as the supervised topics. Using such intermediate "attributes" to aid visual classification has become a popular approach in computer vision [25,24]. After extracting SIFT features [27] from the raw images, quantization into 250 clusters is performed, defining the vocabulary for the bag of visual words. Images with less than two attributes were discarded. The resulting dataset of size 2,275 was equally split into training and test data.

**ACM Conference Data.** The text corpus consists of conference paper abstracts from two groups of conferences. The first group has four conferences related to data mining – WWW, SIGIR, KDD, and ICML, and the second group consists of two VLSI conferences – ISPD and DAC. The classification task is to determine the conference at which the abstract was published. As supervised topics, we use keywords provided by the authors, which are presumably useful in determining the conference venue. Since authors usually take great care in choosing keywords so that their paper is retrieved by relevant searches, we believed that such keywords made a good choice of supervised topics. Part of the data, crawled from ACM's website, was used in [37]. A total of 2,300 abstracts were collected each of which had at least three keywords and an average of 78 ($\pm 33.5$) words. After stop-word removal, the vocabulary size for the assembled data is 13,412 words. The final number of supervised topics, after some standard pre-processing of keywords, is 55. The resulting dataset was equally split into training and test data.

### 5.2 Methodology

In order to demonstrate the contribution of each aspect of the overall model, DSLDA and NP-DSLDA are compared against the following simplified models:

[2] `http://vision.cs.uiuc.edu/attributes/`

- MedLDA with **o**ne-**v**s-**a**ll classification (MedLDA-OVA): A separate model is trained for each class using a one-vs-all approach leaving no possibility of transfer across classes.
- MedLDA with **m**ulti**t**ask **l**earning (MedLDA-MTL): A single model is learned for all classes where the latent topics are shared across classes.
- DSLDA with **o**nly **s**hared **s**upervised **t**opics (DSLDA-OSST): A model in which supervised topics are used and shared across classes but there are no latent topics.
- DSLDA with **n**o **s**hared **l**atent **t**opics (DSLDA-NSLT): A model in which only supervised topics are shared across classes and a separate set of latent topics is maintained for each class.
- **M**ajority **c**lass **m**ethod (MCM): A simple baseline which always picks the most common class in the training data.

These baselines are useful for demonstrating the utility of *both* supervised and latent shared topics for multitask learning in DSLDA. MedLDA-OVA is a non-transfer method, where a separate model is learned for each of the classes, *i.e.* one of the many classes is considered as the positive class and the union of the remaining ones is treated as the negative class. Since the models for each class are trained separately, there is no possibility of sharing inductive information across classes. MedLDA-MTL trains on examples from all classes simultaneously, and thus allows for sharing of inductive information *only* through a common set of latent topics. In DSLDA-OSST, only supervised topics are maintained and knowledge transfer can *only* take place via these supervised topics. DSLDA-NSLT uses shared supervised topics but also includes latent topics which are *not* shared across classes. This model provides for transfer *only* through shared supervised topics but provides extra modeling capacity compared to DSLDA-OSST through the use of latent topics that are not shared. DSLDA and NP-DSLDA are MTL frameworks where both supervised *and* latent topics are shared across all classes. Note that, all of the baselines can be implemented using DSLDA with a proper choice of $\boldsymbol{\Lambda}$ and $\epsilon$. For example, DSLDA-OSST is just a special case of DSLDA with $\epsilon$ fixed at 1.

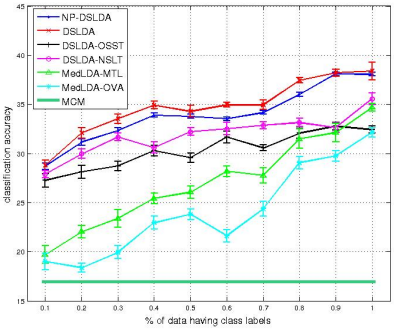


**Fig. 1.** $p_1 = 0.5$ (aYahoo)

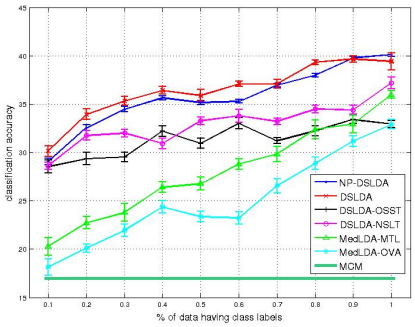


**Fig. 2.** $p_1 = 0.7$ (aYahoo)

**Table 1.** Illustration of Latent and Supervised Topics

| | |
|---|---|
| LT1 | function, label, graph, classification, database, propagation, algorithm, accuracy, minimization, transduction |
| LT2 | performance, design, processor, layer, technology, device, bandwidth, architecture, stack, system |
| CAD | design, optimization, mapping, pin, simulation, cache, programming, routing, biochip, electrode |
| VLSI | design, physical, lithography, optimization, interdependence, global, robust, cells, layout, growth |
| IR | algorithm, web, linear, query, precision, document, repair, site, search, semantics |
| Ranking | integration, catalog, hierarchical, dragpushing, structure, source, sequence, alignment, transfer, flattened, speedup |
| Learning | model, information, trajectory, bandit, mixture, autonomous, hierarchical, feedback, supervised, task |

In order to explore the effect of different amounts of both types of supervision, we varied the amount of both topic-level and class-level supervision. Specifically, we provided topic supervision for a fraction, $p_1$, of the overall training set, and then provided class supervision for only a further fraction $p_2$ of this data. Therefore, only $p_1 * p_2$ of the overall training data has class supervision. By varying the number of latent topics from 20 to 200 in steps of 10, we found that $K_1 = 100$ generally worked the best for all the parametric models. Therefore, we show parametric results for 100 latent topics. For each combination of $(p_1, p_2)$, 50 random trials were performed with $C = 10$. To maintain equal representational capacity, the total number of topics $K$ is held the same across all parametric models (except for DSLDA-OSST where the total number of topics is $K_2$). For NP-DSLDA, following the suggestion of [36], we set $K_1 = 150$ and $T = 40$, which produced uniformly good results. When required, $\epsilon$ was chosen using 5-fold internal cross-validation using the training data.

### 5.3 Results

Figs. 1 and 2 present representative learning curves for the image data, showing how classification accuracy improves as the amount of class supervision ($p_2$) is increased. Results are shown for two different amounts of topic supervision ($p_1 = 0.5$ and $p_1 = 0.7$). Figs. 3 and 4 present similar learning curves for the text data. The error bars in the curves show standard deviations across the 50 trials.

The results demonstrate that DSLDA and NP-DSLDA quite consistently outperform all of the baselines, clearly demonstrating the advantage of combining both types of

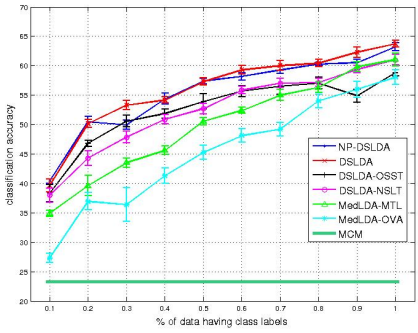


**Fig. 3.** $p_1 = 0.5$ (Conference)

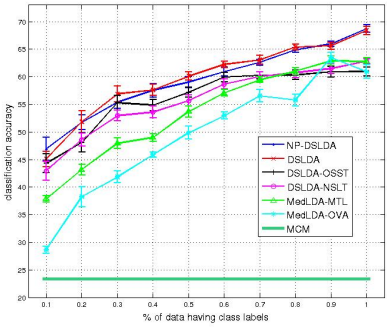


**Fig. 4.** $p_1 = 0.7$ (Conference)

topics. NP-DSLDA performs about as well as DSLDA, for which the optimal number of latent topics has been chosen using an expensive model-selection search. This demonstrates that NP-DSLDA is doing a good job of automatically selecting an appropriate number of latent topics.

Overall, DSLDA-OSST and MedLDA-MTL perform about the same, showing that, individually, both latent and supervised shared topics each support multitask learning about equally well when used alone. However, combining both types of topics provides a clear improvement.

MedLDA-OVA performs quite poorly when there is only a small amount of class supervision (note that this baseline uses *only* class labels). However, the performance approaches the others as the amount of class supervision increases. This is consistent with the intuition that multitask learning is most beneficial when each task has limited supervision and therefore has more to gain by sharing information with other tasks.

Shared supervised topics clearly increase classification accuracy when class supervision is limited (i.e. small values of $p_2$), as shown by the performance of both DSLDA-NSLT and DSLDA-OSST. When $p_2 = 1$ (equal amounts of topic and class supervision), DSLDA-OSST, MedLDA-MTL and MedLDA-OVA all perform similarly; however, by exploiting *both* types of supervision, DSLDA and NP-DSLDA still maintain a performance advantage.

### 5.4 Topic Illustration

In Table 1, we show the most indicative words for several topics discovered by DSLDA from the text data (with $p_1 = 0.8$ and $p_2 = 1$). LT1 and LT2 correspond to the most frequent latent topics assigned to documents in the two broad categories of conferences (data mining and VLSI, respectively). The other five topics are supervised ones. CAD and IR stand for Computer Aided Design and Information Retrieval respectively. The illustrated topics are particularly discriminative when classifying documents.

### 5.5 Discussion

DSLDA-NSLT only allows sharing of supervised topics and its implementation is not straightforward. Since MedLDA-OVA, MedLDA-MTL and DSLDA use $K$ topics (latent or a combination of supervised and latent), to make the comparison fair, it is necessary to maintain the same number of topics for DSLDA-NSLT. This ensures that the models compared have the same representational capacity. Therefore, for each class in DSLDA-NSLT, $k_2/Y$ latent topics are maintained. While training DSLDA-NSLT with examples from the $y^{\text{th}}$ class, only a subset of the first $k_1$ topics (or a subset of the supervised ones based on which of them are present in the training documents) and the next $\left(\frac{(y-1)k_2}{Y} + 1\right)^{\text{th}}$ to $\left(\frac{yk_2}{Y}\right)^{\text{th}}$ topics are considered to be "active" among the latent topics. The other latent topics are assumed to have zero contribution, implying that the parameters associated with these topics are not updated based on observations of documents belonging to class $y$. During testing, however, one needs to project a document onto the entire $K$-dimensional space, and the class label is predicted based on this representation and the parameters $\boldsymbol{r}$.

Overall, the results support the hypothesis that DSLDA's ability to incorporate both supervised and latent topics allow it to achieve better predictive performance compared to baselines that exploit only one, the other, or neither. Furthermore, NP-DSLDA is able to automate model-selection, performing nearly as well as DSLDA with optimally chosen parameters.

## 6 Future Work and Conclusion

This paper has introduced Doubly Supervised LDA (DSLDA) and non-parametric DSLDA (NP-DSLDA), novel approaches that combine the following – generative and discriminative models, latent and supervised topics, and class and topic level supervision, in a principled probabilistic manner. Four ablations of this model are also evaluated in order to understand the individual effects of latent/supervised topics and multitask learning on the overall model performance. The general idea of "double supervision" could be applied to many other models, for example, in multi-layer perceptrons, latent SVMs [40] or in deep belief networks [18]. In MTL, sharing tasks blindly is not always a good approach and further extension with clustered MTL [42] is possible. Based on a very recent study [22], a sampling based algorithm could also be developed for NP-DSLDA, possibly leading to even better performance.

**Acknowledgments.** This research was partially supported by ONR ATL Grant N00014-11-1-0105, NSF Grants (IIS-0713142 and IIS-1016614) and by the Brazilian Research Agencies FAPESP and CNPq.

## References

1. Ando, R., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6, 1817–1853 (2005)
2. Ando, R.K.: Applying alternating structure optimization to word sense disambiguation. In: Proceedings of Computational Natural Language Learning (2006)
3. Argyriou, A., Micchelli, C.A., Pontil, M., Ying, Y.: A spectral regularization framework for multi-task structure learning. In: Proceedings of Neural Information Processing Systsems (2007)
4. Bakker, B., Heskes, T.: Task clustering and gating for Bayesian multitask learning. Journal of Machine Learning Research 4 (2003)
5. Ben-David, S., Schuller, R.: Exploiting task relatedness for multiple task learning. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 567–580. Springer, Heidelberg (2003)
6. Bickel, S., Bogojeska, J., Lengauer, T., Scheffer, T.: Multi-task learning for HIV therapy screening. In: Proceedings of International Conference on Machine Learning, pp. 56–63. ACM, New York (2008)
7. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review 94, 115–147 (1987)
8. Blei, D.M., Mcauliffe, J.D.: Supervised topic models. In: Proceedings of Neural Information Processing Systems (2007)

9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
10. Bollacker, K.D., Ghosh, J.: Knowledge transfer mechanisms for characterizing image datasets. In: Soft Computing and Image Processing. Physica-Verlag, Heidelberg (2000)
11. Caruana, R.: Multitask learning. Machine Learning 28, 41–75 (1997)
12. Chang, J., Blei, D.: Relational topic models for document networks. In: Proceedings of Artificial Intelligence and Statistics (2009)
13. Eisenstein, J., Ahmed, A., Xing, E.P.: Sparse additive generative models of text. In: Proceedings of International Conference on Machine Learning, pp. 1041–1048 (2011)
14. Evgeniou, T., Micchelli, C.A., Pontil, M.: Learning multiple tasks with kernel methods. Journal of Machine Learning Research 6, 615–637 (2005)
15. Evgeniou, T., Pontil, M., Toubia, O.: A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. Marketing Science 26(6), 805–818 (2007)
16. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
17. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proceedings of Computer Vision and Pattern Recognition (2009)
18. Hinton, G.E., Osindero, S.: A fast learning algorithm for deep belief nets. Neural Computation 18, 2006 (2006)
19. Jacob, L., Bach, F., Vert, J.-P.: Clustered multi-task learning: A convex formulation. CoRR, abs/0809.2085 (2008)
20. Jalali, A., Ravikumar, P., Sanghavi, S., Ruan, C.: A Dirty Model for Multi-task Learning. In: Proceedings of Neural Information Processing Systems (December 2010)
21. Jenatton, R., Audibert, J., Bach, F.: Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research 12, 2777–2824 (2011)
22. Jiang, Q., Zhu, J., Sun, M., Xing, E.: Monte carlo methods for maximum margin supervised topic models. In: Proceedings of Neural Information Processing Systems, pp. 1601–1609 (2012)
23. Kim, S., Xing, E.P.: Tree-guided group lasso for multi-task regression with structured sparsity. In: Proceedings of International Conference on Machine Learning, pp. 543–550 (2010)
24. Kovashka, A., Vijayanarasimhan, S., Grauman, K.: Actively selecting annotations among objects and attributes. In: International Conference on Computer Vision, pp. 1403–1410. IEEE (2011)
25. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by betweenclass attribute transfer. In: Proceedings of Computer Vision and Pattern Recognition (2009)
26. Low, Y., Agarwal, D., Smola, A.J.: Multiple domain user personalization. In: Proceedings of Knowledge Discovery and Data Mining, pp. 123–131 (2011)
27. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
28. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359 (2010)
29. Passos, A., Rai, P., Wainer, J., Daumé III, H.: Flexible modeling of latent task structures in multitask learning. In: Proceedings of International Conference on Machine Learning (2012)
30. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T., Csail, M.: Hidden-state conditional random fields. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2007)
31. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of Empirical Methods in Natural Language Processing, pp. 248–256 (2009)

32. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. CoRR, abs/1107.2462 (2011)
33. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101, 1566–1581 (2006)
34. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE Trans. Pattern Anal. Mach. Intell. 29(5), 854–869 (2007)
35. Wang, C., Blei, D.M., Li, F.F.: Simultaneous image classification and annotation. In: Proceedings of Computer Vision and Pattern Recognition, pp. 1903–1910 (2009)
36. Wang, C., Paisley, J.W., Blei, D.M.: Online variational inference for the hierarchical Dirichlet process. Journal of Machine Learning Research - Proceedings Track 15, 752–760 (2011)
37. Wang, C., Thiesson, B., Meek, C., Blei, D.: Markov topic models. In: Proceedings of Artificial Intelligence and Statistics (2009)
38. Weinberger, K., Dasgupta, A., Langford, J., Smola, A., Attenberg, J.: Feature hashing for large scale multitask learning. In: Proceedings of International Conference on Machine Learning, pp. 1113–1120 (2009)
39. Xue, Y., Liao, X., Carin, L., Krishnapuram, B.: Multi-task learning for classification with Dirichlet process priors. Journal of Machine Learning Research 8, 35–63 (2007)
40. Yu, C.J., Joachims, T.: Learning structural SVMs with latent variables. In: Proceedings of International Conference on Machine Learning, pp. 1169–1176 (2009)
41. Zhang, J., Ghahramani, Z., Yang, Y.: Flexible latent variable models for multi-task learning. Machine Learning 73(3), 221–242 (2008)
42. Zhou, J., Chen, J., Ye, J.: Clustered Multi-Task Learning Via Alternating Structure Optimization. In: Proceedings of Neural Information Processing Systems (2011)
43. Zhu, J., Ahmed, A., Xing, E.P.: MedLDA: maximum margin supervised topic models for regression and classification. In: Proceedings of International Conference on Machine Learning, pp. 1257–1264 (2009)