In Proceedings of the 2014 SIAM International Conference on Data Mining (SDM14), Philadelphia, Pennsylvania, April 2014.

Active Multitask Learning Using Both Latent and Supervised Shared Topics

Ayan Acharya* Raymond J. Mooney* Joydeep Ghosh*

Abstract

Multitask learning (MTL) *via* a shared representation has been adopted to alleviate problems with sparsity of labeled data across different learning tasks. Active learning, on the other hand, reduces the cost of labeling examples by making informative queries over an unlabeled pool of data. Therefore, a unification of both of these approaches can potentially be useful in settings where labeled information is expensive to obtain but the learning tasks or domains have some common characteristics. This paper introduces two such models – Active Doubly Supervised Latent Dirichlet Allocation (Act-DSLDA) and its non-parametric variation (Act-NPDSLDA) that integrate MTL and active learning in the same framework. These models make use of both latent and supervised shared topics to accomplish multitask learning. Experimental results on both document and image classification show that integrating MTL and active learning along with shared latent and supervised topics is superior to other methods which do not employ all of these components.

Keywords

Active Learning, Multitask Learning, Topic Model.

1 Introduction

Building an automated object detector in computer vision is often challenging. Object categories abound in nature and it is expensive to obtain sufficient labeled examples for all of them. Computer vision researchers have attempted to overcome this challenge by either gathering large datasets of web images [11, 14, 35] or by formulating new methods that reduce the amount of human supervision required.

One such method, partly inspired by human perception and learning from high-level object descriptions, utilizes *attributes* which describe abstract object properties shared by many categories [13, 22, 21, 1]. These attributes serve as an intermediate layer in a classifier cascade. If the *shared* attributes transcend object class boundaries, such a classifier cascade is beneficial for *transfer learning* [26] where fewer labeled examples are available for some object categories compared to others [22]. For example, in the aYahoo image dataset [13] used in our experiments, there are 12 classes, including carriage, donkey, goat, and zebra. Each image is also annotated for 64 relevant visual attributes, such as "has head" and "has wheel." Learning to recognize such attributes improves classification across multiple related classes. Another well-known approach to reducing supervision is *active learning*, where a system can request labels for the most informative training examples [28, 16, 18, 21].

In this paper, our objective is to combine these two orthogonal approaches in order to leverage the benefits of both – learning from a shared abstract feature space and making active queries. In particular, we build on a recent approach proposed in [1] where multitask learning (MTL) [6] is accomplished using both shared supervised attributes and a shared latent (i.e. unsupervised) set of features. 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. This approach is in contrast to "isolated" training of tasks where each task is learned independently using a separate model.

The paper is organized as follows. We present related work in Section 2, followed by the descriptions of two of our models Active Doubly Supervised Latent Dirichlet Allocation (Act-DSLDA) and a non-parametric variation of the same (Act-NPDSLDA) in Sections 3 and 4 respectively. Experimental results on both multi-class image and document categorization are presented in Section 5. 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 Background and Related Work

2.1 Statistical Topic Models LDA [3] 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 multi-

*University of Texas at Austin, Austin, TX, USA. Email: {aacharya@ece, mooney@cs, ghosh@ece}.utexas.edu

ple topics. The unsupervised LDA has been extended to account for supervision by labeling each document with its set of topics [31, 34]. In *Labeled LDA* (LLDA [31]), the primary objective is to build a model of the words that indicate the presence of certain topic labels. Some other researchers [2, 44, 9] assume that supervision is provided for a single *response variable* to be predicted for a given document. In *Maximum Entropy Discriminative LDA* (MedLDA) [44], 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.

2.2 Active Learning *via* Expected Error Reduction Of the several measures for selecting labels in active learning algorithms, a decision-theoretic approach called Expected Error Reduction [33] has been used quite extensively in practice [21, 37]. This approach aims to measure how much the generalization error of a model is likely to be reduced based on some labeled information y of an instance $\boldsymbol{x}$ taken from the unlabeled pool $\mathcal{U}$. The idea is to estimate the expected future error of a model trained using $\mathcal{L} \cup \langle \boldsymbol{x}, y \rangle$ on the remaining unlabeled instances in $\mathcal{U}$, and query the instance with minimal expected future error. Here $\mathcal{L}$ denotes the labeled pool of data. One approach is to minimize the expected 0/1 loss:

$$\boldsymbol{x}^*_{0/1} = \operatorname*{argmax}_{\boldsymbol{x}} \sum_n P_{\boldsymbol{\kappa}}(y_n|\boldsymbol{x}) \left(\sum_{u=1}^{U} 1 - P_{\boldsymbol{\kappa}^{+\langle \boldsymbol{x}, y_n \rangle}} \left(\hat{y}, \boldsymbol{x}^{(u)} \right) \right). \tag{2.1}$$

where $\boldsymbol{\kappa}^{+\langle \boldsymbol{x}, y_n \rangle}$ refers to the new model after it has been re-trained with the training set $\mathcal{L} \cup \langle \boldsymbol{x}, y_n \rangle$. Note that we do not know the true label for each query instance, so we approximate using expectation over all possible labels under the current model. The objective is to reduce the expected number of incorrect predictions.

2.3 Active Knowledge Transfer There has been some effort to integrate active and transfer learning in the same framework. In [19] the authors utilized a maximum likelihood classifier to learn parameters from the source domain and use these parameters to seed the EM algorithm that explains the unlabeled data in the target domain. The example which contributed to maximum expected KL divergence of the posterior distribution with the prior distribution was selected in the active step. In [30], the source data is first used to train a classifier, the parameters of which are later updated in an online manner with new examples actively selected from the target domain. The active selection criterion is based on uncertainty sampling [37]. Similarly, in [8], a naïve Bayes classifier is first trained with examples from the source domain and then incrementally updated with data from the target domain selected using uncertainty sampling. The method proposed in [38] maintains a classifier trained on the source domain(s) and the prediction of this classifier is trusted only when the likelihood of the data in the target domain is sufficiently high. In case of lower likelihood, domain experts are asked to label the example. Harpale & Young [15] proposed active multitask learning for adaptive filtering [32] where the underlying classifier is logistic regression with Dirichlet process priors. Any feedback provided in the active selection phase improves both the task-specific and the global performance *via* a measure called *utility gain* [15]. Saha *et al.* [36] formulated an online active multitask learning framework where the information provided for one task is utilized for other tasks through a task correlation matrix. The updates are similar to perceptron updates. For active selection, they use a margin based sampling scheme which is a modified version of the sampling scheme used in [7].

In contrast to this previous work, our approach employs a topic-modeling framework and uses expected error reduction for active selection. Such an active selection mechanism necessitates fast incremental update of model parameters, and hence the inference and estimation problems become challenging. This approach to active selection is more immune to noisy observations compared to simpler methods such as uncertainty sampling [37]. Additionally, our approach can query both class labels *and* supervised topics (i.e. attributes), which has not previously been explored in the context of MTL.

2.4 Multitask Learning Using Both Shared Latent and Supervised Topics In multitask learning (MTL [6]), a single model is simultaneously trained to perform multiple related tasks. Many different MTL approaches have been proposed over the past 15 years (*e.g.*, see [41, 26, 27] and references therein). These include different learning methods, such as empirical risk minimization using group-sparse regularizers [20, 17], hierarchical Bayesian models [43, 23] and hidden conditional random fields [29]. In an MTL framework, if the tasks are related, training one task should provide helpful "inductive bias" for learning the other tasks.

In particular, Acharya *et al.* [1] proposed two models – **D**oubly **S**upervised **L**atent **D**irichlet **A**llocation (DSLDA) and its **n**on-**p**arametric counterpart (NPDSLDA) which support the prediction of multiple response variables based on a combination of both supervised and latent topics. In computer vision terminology, the supervised topics correspond to attributes provided by human experts. In both text and vision domains, Acharya *et al.* [1] showed that incorporating

both supervised and latent topics achieves better predictive performance compared to baselines that exploit only one, the other, or neither. In our paper, we extend these models to include *active* sample selection. This extension is non-trivial and requires several modifications to the inference and learning methods. With that objective in mind, the next two sub-sections discuss the incremental EM algorithm and the online support vector machine used to adapt DSLDA.

2.5 Incremental EM Algorithm The EM algorithm proposed by Dempster *et al.* [10] can be viewed as a joint maximization problem over $q(.)$, the conditional distribution of the hidden variables $\boldsymbol{Z}$ given the model parameters $\boldsymbol{\kappa}$ and the observed variables $\boldsymbol{X}$. The relevant objective function is given as follows:

$$F(q, \boldsymbol{\kappa}) = \mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\kappa}))] + H(q), \tag{2.2}$$

where $H(q)$ is the entropy of the distribution $q(.)$. Often, $q(.)$ is restricted to a family of distributions $\mathcal{Q}$. It can be shown that if $\boldsymbol{\theta}^*$ is the maximizer of the above objective F then it also maximizes the likelihood of the observed data. In most of the models used in practice, the joint distribution is assumed to factorize over the instances implying that $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\kappa}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\kappa})$. One can further restrict the family of distributions $\mathcal{Q}$ to maximize over in Eq. (2.2) to the factorized form: $q(\boldsymbol{Z}) = \prod_{n=1}^{N} q(\boldsymbol{z}_n|\boldsymbol{x}_n) = \prod_{n=1}^{N} q_n$.

An incremental variant of the EM algorithm that exploits such separability structure in both $p(.)$ and $q(.)$ was first proposed by Neal & Hinton [25]. Under such structure, the objective function in Eq. (2.2) decomposes over the observations $F(q, \boldsymbol{\theta}) = \sum_{n=1}^{N} F_n(q_n, \boldsymbol{\kappa})$, and the following incremental algorithm can instead be used to maximize F:

- **E step:** Choose some observation n to be updated over, set $q_{n'}^{(t)} = q_{n'}^{(t-1)}$ for $n' \neq n$ (no update) and set $q_n^{(t)} = \underset{q_n}{\operatorname{argmax}}\ F_n(q_n, \boldsymbol{\kappa}^{(t-1)})$.
- **M step:** $\boldsymbol{\kappa}^{(t)} = \underset{\boldsymbol{\kappa}}{\operatorname{argmax}}\ F(q^{(t)}, \boldsymbol{\kappa})$.

2.6 Online Support Vector Machines The online SVM proposed by Bordes *et al.* [4, 5] has three distinct modules that work in unison to provide a scalable learning mechanism. These modules are named "ProcessNew", "ProcessOld" and "Optimize". All of these modules use a common operation called "SMOStep" and the memory footprint is limited to the support vectors and associated gradient information. The module "ProcessNew" operates on a pattern that is not a support pattern. In such an update, one of the classes is chosen as the label of the support pattern and the other class is chosen such that it defines a feasible direction with the highest gradient. It then performs an SMO step with the example and the selected classes. The module "ProcessOld" randomly picks a support pattern and chooses two classes that define the feasible direction with the highest gradient for that support pattern. "Optimize" resembles "ProcessOld" but picks two classes among those that correspond to existing support vectors.

3 Active Doubly Supervised Latent Dirichlet Allocation (Act-DSLDA)

We will treat examples as "documents" which consist of a "bag of words" for text or a "bag of visual words" for images. Assume we are given an initial training corpus $\mathcal{L}$ with N documents belonging to Y different classes. Further assume that each of these training documents is also annotated with a set of K_2 different "supervised topics". The objective is to train a model using the words in a document, as well as the associated supervised topics and class labels, and then use this model to classify completely unlabeled test documents for which no topics or class labels are provided.

When the learning starts, $\mathcal{L}$ is assumed to have fully labeled documents. However, as the learning progresses more documents are added to the pool $\mathcal{L}$ with class and/or a subset of supervised topics labeled. Therefore, at any intermediate point of the learning process, $\mathcal{L}$ can be assumed to contain several sets: $\mathcal{L} = \{\mathcal{T} \cup \mathcal{T}_C \cup \mathcal{T}_{A_1} \cup \mathcal{T}_{A_2} \cup \cdots \cup \mathcal{T}_{A_{K_2}}\}$, where $\mathcal{T}$ contains fully labeled documents (*i.e.* with class and all supervised topics labeled), $\mathcal{T}_C$ are the documents that have class labels, and $1 \leq k \leq K_2$, $\mathcal{T}_{A_k}$ are the documents that have the k^{th} supervised topic labeled. Since, human-provided labels are expensive to obtain, we design an active learning framework where the model can query over an unlabeled pool $\mathcal{U}$ and request either class labels or a subset of the supervised topics.

Please note that the proposed frameworks support general MTL; however, our datasets, as explained in Section 5, happen to be multiclass, where each class is treated as a separate "task" (as typical in multiclass learning based on binary classifiers). However, the frameworks are not in any way restricted to multiclass MTL. The Act-DSLDA generative model is defined as follows. For the n^{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, 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 ($K = K_1 + K_2$). Latent topics are never observed, while supervised topics are observed in the training data but not in the 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. $\mathbf{\Lambda}_n$ is a diagonal binary matrix of dimension $K \times K$. The k^{th} diagonal entry is unity if *either* $1 \leq k \leq K_1$ *or* $K_1 < k \leq K$ and the n^{th} document is tagged with the k^{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.

In the test data, the supervised topics are not observed and one has to infer them from either the parameters of the model or use some other auxiliary information. Since one of our objectives is to query over the supervised topics as well as the final category, we train a set of binary SVM classifiers that can predict the individual attributes from the features of the data. We denote the parameters of such classifiers by $\{\boldsymbol{r}_{2k}\}_{1 \leq k \leq K_2}$. This is important to get an uncertainty measure over the supervised topics. To further clarify the issue, let us consider that only one supervised topic has to be labeled by the annotator for the n^{th} document from the set of supervised topics of size K_2. To select the most uncertain topic, one needs to compare the uncertainty of predicting the presence or absence of the individual topics. This uncertainty is different from that calculated from the conditional distribution calculated from the posterior over θ_n.

For the m^{th} word in the n^{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 ϵ 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^{th} topic.

For the n^{th} document, generate $Y_n = \arg\max_y \boldsymbol{r}_{1y}^T \mathbb{E}(\bar{\mathbf{z}}_n)$ where Y_n is the class label associated with the n^{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. $\boldsymbol{r}_{1y}$ is a K-dimensional real vector corresponding to the y^{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^{th} document. The maximization problem to generate Y_n (i.e. the classification problem) is carried out using the max-margin principle and we use online SVMs [4, 5] for such updates. Since the model has to be updated incrementally in the active selection step, a batch SVM solver is not applicable, while an online SVM allows one to update the learned weights incrementally given each new example. Note that predicting each class is treated as a separate task, and that the shared topics are useful for generalizing the performance of the model across classes.

3.1 Inference and Learning Inference and parameter estimation have two phases – one for the batch case when the model is trained with fully labeled data, and the other for the active selection step where the model has to be incrementally updated to observe the effect of any labeled information that is queried from the oracle.

3.1.1 Learning in Batch Mode 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:

$$p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\kappa}_0) = \prod_{n=1}^{N} p(\boldsymbol{\theta}_n|\boldsymbol{\alpha}_n) \prod_{m=1}^{M_n} p(z_{nm}|\boldsymbol{\theta}'_n) p(w_{nm}|\boldsymbol{\beta}_{\boldsymbol{z}_{nm}}). \tag{3.3}$$

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

$$q(\boldsymbol{Z}|\{\boldsymbol{\kappa}_n\}_{n=1}^{N}) = \prod_{n=1}^{N} q(\boldsymbol{\theta}_n|\boldsymbol{\gamma}_n) \prod_{m=1}^{M_n} q(z_{nm}|\boldsymbol{\phi}_{nm}), \tag{3.4}$$

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^{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, Act-DSLDA reduces to solving the following optimization problem[1]:

$$\min_{q, \boldsymbol{\kappa}_0, \{\xi_n\}} \frac{1}{2}||\boldsymbol{r}_1||^2 - \mathcal{L}(q(\boldsymbol{Z})) + C \sum_{n=1}^{N} \xi_n \mathbb{I}_{\mathcal{T}_C, n},$$

$$\text{s.t. } \forall n \in \mathcal{T}_C, y \neq Y_n : \mathbb{E}[\boldsymbol{r}_1^T \Delta f_n(y)] \geq 1 - \xi_n; \xi_n \geq 0. \tag{3.5}$$

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. $\mathbb{E}[\boldsymbol{r}_1^T \Delta f_n(y)]$ is

[1] Please see [44] for further details.

the "expected margin" over which the true label Y_n is preferred over a prediction y. From this viewpoint, Act-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. The indicator variable $\mathbb{I}_{\mathcal{T}_C,n}$ is unity if the n^{th} document has a class label (*i.e.* $n \in \mathcal{T}_C$) and 0 otherwise. This implies that only the documents that have class labels are used to update the parameters of the online SVM.

Let $\mathcal{Q}$ be the set of all distributions having a fully factorized form as given in (3.4). Note that such a factorized approximation makes the use of incremental variation of EM possible in the active selection step following the discussion in Section 2.5. Let the distribution q^* from the set $\mathcal{Q}$ optimize the objective in Eq. (3.5). The optimal values of the corresponding variational parameters are same as those of DSLDA [1]. 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.

During testing, one does not observe a document's supervised topics and instead an approximate solution, as also used in [31, 1], is employed where the variables $\{\boldsymbol{\Lambda}_n\}$ are assumed to be absent altogether in the test phase, and the problem is treated as inference in MedLDA with K latent topics.

In the M step, the objective in Eq. (3.5) is maximized w.r.t $\boldsymbol{\kappa}_0$. The optimal value of β_{kv} is again similar to that of DSLDA [1]. However, numerical methods for optimization are required to update $\boldsymbol{\alpha}_1$ or $\boldsymbol{\alpha}_2$. The update for the parameters $\{\boldsymbol{r}_{1y}\}_{y=1}^{Y}$ is carried out using online SVM [4, 5] following Eq. (3.5).

3.1.2 Incremental Learning in Active Selection

The method of Expected Entropy Reduction requires one to take an example from the unlabeled pool and one of its possible labels, update the model, and observe the generalized error on the unlabeled pool. This process is computationally expensive unless there is an efficient way to update the model incrementally. The incremental view of EM and the online SVM framework are appropriate for such updates.

Consider that a completely unlabeled or partially labeled document, indexed by n', is to be included in the labeled pool with one of the $(K_2 + 1)$ labels (one for the class label and each different supervised topic), indexed by k'. In the E step, variational parameters corresponding to all other documents except for the n'th one are kept fixed and the variational parameters for only the n'th document are updated. In the M-step, we keep the priors $\{\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}\}$ over the topics and the SVM parameters $\boldsymbol{r}_2$ fixed as there is no easy way to update such parameters incrementally. From the empirical point of view, these parameters do not change much w.r.t. the variational parameters (or features in topic space representation) of a single document. However, the update of the parameters $\{\boldsymbol{\beta}, \boldsymbol{r}_1\}$ is easier. Updating $\boldsymbol{\beta}$ is accomplished by a simple update of the sufficient statistics. Updating $\boldsymbol{r}_1$ is done using the "ProcessNew" operation of online SVM followed by a few iterations of "ProcessOld". The selection of the document-label pair is guided by the measure given in Eq. (2.1). Note that since SVM uses hinge loss which, in turn, upper bounds the 0–1 loss in classification, use of the measure from Eq. (2.1) for active query selection is justified.

From the modeling perspective, the difference between DSLDA [1] and Act-DSLDA lies in maintaining attribute classifiers and ignoring documents in the max-margin learning that do not have any class label. Online SVM for max-margin learning is essential in the batch mode just to maintain the support vectors and incrementally update them in the active selection step. One could also use incremental EM for batch mode training. However, that is computationally more complex when the labeled dataset is large, as the E step for each document is followed by an M-step in incremental EM.

4 Active Non-parametric DSLDA (Act-NPDSLDA)

A non-parametric extension of Act-DSLDA (Act-NPDSLDA) automatically determines the best number of latent topics for modeling the given data. It uses a modified stick breaking construction of Hierarchical Dirichlet Process (HDP), recently introduced in [40], to make variational inference feasible. The Act-NPDSLDA generative model is 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. Also, sample $\beta'_{k_1} \sim \text{Beta}(1, \delta_0)\ \forall k_1 \in \{1, 2, \cdots, \infty\}$.

• For the n^{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^{th} diagonal entry is unity if the n^{th} word is tagged with the k^{th} supervised topic. Similar to the case of Act-DSLDA, in the test data, the supervised topics are not observed and the set of binary SVM classifiers, trained with document-attribute pair data, are used to predict the individual attributes from the input features. The parameters of such classifiers are denoted by $\{\boldsymbol{r}_{2k}\}_{1 \le k \le K_2}$.

• $\forall n, \forall t \in \{1, 2, \cdots, \infty\}$, 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 $\boldsymbol{c}_n$.

• For the m^{th} word in the n^{th} document, sample $z_{nm} \sim \text{multinomial}((1-\epsilon)\boldsymbol{\pi}_n^{(1)}, \epsilon\boldsymbol{\pi}_n^{(2)})$. This implies that with probability ϵ, a topic is selected from the set of supervised topics and with probability $(1-\epsilon)$, a topic is chosen from the set of unsupervised topics. Sample w_{nm} from a multinomial given by Eq. (3).

• For the n^{th} document, generate $Y_n = \arg\max_y \boldsymbol{r}_{1y}^T \mathbb{E}(\bar{\mathbf{z}}_n)$ where Y_n is the class label associated with the n^{th} document, $\bar{\mathbf{z}}_n = \sum_{m=1}^{M_n} \mathbf{z}_{nm}/M_n$. The maximization problem to generate Y_n (i.e. the classification problem) is carried out using an online support vector machine. The joint distribution of the hidden and observed variables is given in Eq. (1).

4.1 Inference and Learning

4.1.1 Learning in Batch Mode As an approximation to the posterior distribution over the hidden variables, we use the factorized distribution given in Eq. (2). $\boldsymbol{\kappa}_0$ and $\boldsymbol{\kappa}$ denote the sets of model and variational parameters, respectively. $\bar{K_1}$ 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 the Dirichlet, each of dimension V. $\{u_{k_1}, v_{k_1}\}$ and $\{a_{nt}, b_{nt}\}$ are the parameters of Beta distribution corresponding to corpus level and document level sticks respectively. $\{\boldsymbol{\varphi}_{nt}\}$ are multinomial parameters of dimension $\bar{K_1}$ and $\{\boldsymbol{\zeta}_{nm}\}$ are multinomials of dimension $(T+K_2)$. $\{\boldsymbol{\gamma}_n\}_n$ are parameters of the Dirichlet distribution of dimension K_2.

The underlying optimization problem takes the same form as in Eq. (3.5). 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 $\bar{K_1}$ number of corpus level sticks, $\bar{\boldsymbol{s}}_n$ turns out to be of dimension $(\bar{K_1}+K_2)$ and the feature vector $f(y, \bar{\boldsymbol{s}}_n)$ constitutes $Y(\bar{K_1}+K_2)$ elements. The components of $f(y, \bar{\boldsymbol{s}}_n)$ from $(y-1)(\bar{K_1}+K_2)+1$ to $y(\bar{K_1}+K_2)$ are those of the vector $\bar{\boldsymbol{s}}_n$ and all the others are 0. The E-step update equations of Act-NPDSLDA are similar to NP-DSLDA [1]. The M-step updates are similar to Act-DSLDA and are omitted here due to space constraints.

4.1.2 Incremental Learning in Active Selection Assume that a completely unlabeled or partially labeled document, indexed by n', is to be included in the labeled pool with the k'th label. In the E step, variational parameters corresponding to all other documents except for the n'th one is kept fixed and the variational parameters for only the n'th document are updated. The incremental update of the "global" variational parameters $\{u_{k_1}, v_{k_1}\}_{k_1=1}^{K_1}$ is also straightforward following the equations given in [1]. In the M-step, we keep the priors $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\alpha}^{(2)}\}$ and the SVM parameters $\boldsymbol{r}_2$ fixed but the parameters $\boldsymbol{r}_1$ are updated using online SVM.

5 Experimental Evaluation

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

5.1.1 aYahoo Data The first set of experiments was conducted with the aYahoo image dataset from [13] which has 12 classes – carriage, centaur, bag, building, donkey, goat, jetski, monkey, mug, statue, wolf, and zebra. Each image is annotated with 64 relevant visual attributes such as "has head", "has wheel", "has torso" and others, which we use as the supervised topics. aYahoo has been used as a benchmark dataset for knowledge transfer with intermediate "attributes" in computer vision [22, 21]. After extracting SIFT features [24] from the raw images, quantization into 250 clusters was performed using K-means clustering algorithm, defining the vocabulary for a "bag of visual words". Images with less than two attributes were discarded. The resulting dataset of 2,275 images was equally split into training and test data.

5.1.2 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. 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 number of supervised topics (i.e. keywords) is 55. The resulting dataset was equally split into training and test data.

5.2 Methodology Act-DSLDA and Act-NPDSLDA are compared against the following simplified models:

• Active Learning in MedLDA with **o**ne-**v**s-**a**ll classification (Act-MedLDA-OVA): A separate MedLDA

Joint Distribution of Act-NPDSLDA

$$p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\kappa}_0) = \prod_{k_1=1}^{\infty} p(\phi_{k_1}|\boldsymbol{\eta}_1)p(\beta'_{k_1}|\boldsymbol{\delta}_0) \prod_{k_2=1}^{K_2} p(\phi_{k_2}|\boldsymbol{\eta}_2) \prod_{n=1}^{N} p(\boldsymbol{\pi}_n^{(2)}|\boldsymbol{\alpha}_2) \prod_{t=1}^{\infty} p(\boldsymbol{\pi}_{nt}^{\prime(1)}|\boldsymbol{\alpha}_0)p(c_{nt}|\boldsymbol{\beta}') \prod_{m=1}^{M_n} p(z_{nm}|\boldsymbol{\pi}_n^{(1)}, \boldsymbol{\pi}_n^{(2)}, \epsilon)p(w_{nm}|\boldsymbol{\phi}, c_{nz_{nm}}, z_{nm}). \quad (1)$$

Variational Distribution of Act-NPDSLDA

$$q(\boldsymbol{Z}|\boldsymbol{\kappa}) = \prod_{k_1=1}^{\bar{K}_1} q(\phi_{k_1}|\boldsymbol{\lambda}_{k_1}) \prod_{k_2=1}^{K_2} q(\phi_{k_2}|\boldsymbol{\lambda}_{k_2}) \prod_{k_1=1}^{\bar{K}_1-1} q(\beta'_{k_1}|u_{k_1}, v_{k_1}) \prod_{n=1}^{N} q(\boldsymbol{\pi}_n^{(2)}|\boldsymbol{\gamma}_n) \prod_{t=1}^{T-1} q(\boldsymbol{\pi}_{nt}^{\prime(1)}|a_{nt}, b_{nt}) \prod_{t=1}^{T} q(c_{nt}|\boldsymbol{\varphi}_{nt}) \prod_{m=1}^{M_n} q(z_{nm}|\boldsymbol{\zeta}_{nm}). \quad (2)$$

Multinomial Distribution for Sampling Words in Act-NPDSLDA

$$\prod_{k_1=1}^{\infty} \prod_{v=1}^{V} \phi_{k_1 v}^{\mathbb{I}\{w_{nm}=v\}\mathbb{I}\{c_{nz_{nm}}=k_1 \in \{1,\cdots,\infty\}\}} \prod_{k_2=1}^{K_2} \prod_{v=1}^{V} \phi_{k_2 v}^{\mathbb{I}\{w_{nm}=v\}\mathbb{I}\{z_{nm}=k_2 \in \{1,\cdots,K_2\}\}}. \quad (3)$$

Table 1: Distributions in Act-NPDSLDA

model is trained for each class using a one-vs-all approach leaving no possibility of transfer across classes. Supervised topics are not included in such modeling and the class labels are also obtained using active learning.

• Active Learning in MedLDA with **m**ulti**t**ask **l**earning (Act-MedLDA-MTL): A single MedLDA model is learned for all classes where the latent topics are shared across classes. Again, supervised topics are not used and the class labels are obtained using active learning. This baseline is intended to be stronger than Act-MedLDA-OVA where the latent topics are not shared.

• Act-DSLDA with **o**nly **s**hared **s**upervised **t**opics (Act-DSLDA-OSST): A model in which supervised topics are used and shared across classes but there are no latent topics. Both the supervised topics and the class labels are queried using active selection strategy.

• Act-DSLDA with **n**o **s**hared **l**atent **t**opics (Act-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. Both the supervised topics and the class labels are queried using active selection strategy. This model has richer representational capacity compared to Act-DSLDA-OSST which does not use any latent topics at all.

• **R**andom **s**election of only **c**lass labels (RSC) – A MedLDA-MTL model where examples with only class labels are selected at random but supervised topics are not used at all. Note that designing a DSLDA based model where only class labels are selected at random is tricky as one needs to balance the number of supervised topics queried and the number of class labels selected at random. This baseline shows the utility of active selection of classes in the MedLDA-MTL framework.

• **R**andom **s**election of **c**lass and **a**ttribute labels (RSCA) – A DSLDA model where both queries for class and the supervised topics are selected at random. This baseline is weaker than RSC since the supervised topics are generally less informative compared to class labels. Both RSC and RSCA are used to exhibit the utility of active learning for both class and supervised topic selection.

5.3 Results In the experiments with both image and text data, we start with a completely labeled dataset $\mathcal{L}$ consisting of 300 documents. In every active iteration, we query for 50 labels (class labels or supervised topics). Figs. 1 and 2 present representative learning curves for the image and the text data respectively, showing how classification accuracy improves as the amount of supervision is increased. The error bars in the curves show standard deviations across 20 trials.

5.4 Discussion Act-DSLDA and Act-NPDSLDA quite consistently outperform all of the baselines, clearly demonstrating the advantage of combining both types of topics and integrating active learning and transfer learning in the same framework. Act-NPDSLDA performs about as well or better as Act-DSLDA, for which the optimal number of latent topics has been chosen using an expensive model-selection search.

As to be expected, the active DSLDA methods' advantage over their random selection counterpart (RSC) is greatest at the lower end of the learning curve. Act-MedLDA-OVA does a little better than RSCA showing that the active selection of class labels helps even if there is no transfer across classes. Act-MedLDA-MTL consistently outperforms Act-MedLDA-OVA as well as RSC showing that active transfer learning is beneficial for MedLDA-MTL. Act-DSLDA-OSST does better than both Act-MedLDA-MTL and RSC towards the lower end of the learning curve but with more labeled information this model does not perform that well since it does not use latent topics. Act-DSLDA-NSLT also performs better than Act-DSLDA-OSST because the former utilizes latent topics.

Figs. 3 and 4 show the percentage (out of 50 queries) of class labels and supervised topics queried by Act-DSLDA at each iteration step in the vision and text data, respectively. Initially, the model queries for

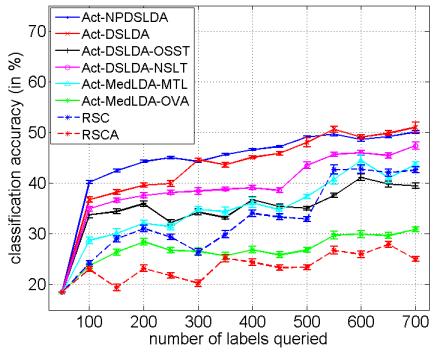

Figure 1: aYahoo Learning Curves

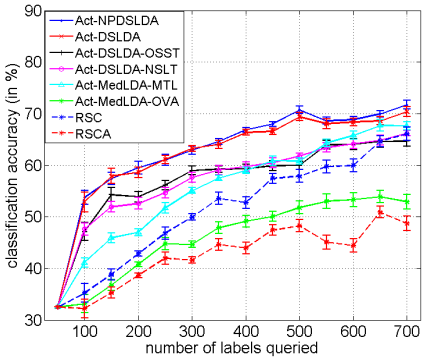

Figure 2: ACM Conference Learning Curves

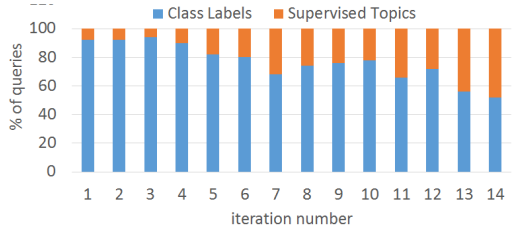

Figure 3: aYahoo Query Distribution

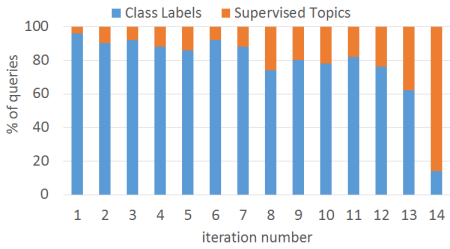

Figure 4: ACM Conference Query Distribution

more class labels but towards the end of the learning curve, more supervised topics are queried. By the 14^{th} iteration, the class labels of all the documents in the training set are queried. From the 15^{th} iteration onwards, only supervised topics are queried. This observation is not that surprising since the class labels are more discriminative compared to the supervised topics and hence are queried more. However, queries of supervised topics are also helpful and allow continued improvement later in the learning curve.

6 Future Work and Conclusion

This paper has introduced two new models for active multitask learning. Experimental results comparing to six different ablations of these models demonstrate the utility of integrating active and multitask learning in one framework that also unifies latent and supervised shared topics. One could additionally actively query for rationales [42, 12] and further improve the predictive performance. The computational complexity of the proposed models largely depends on the active selection mechanism adopted. For large scale applications, one needs to use better approximation techniques for active selection as suggested in [16, 39].

Acknowledgements

This research was partially supported by ONR ATL Grant N00014-11-1-0105 and NSF Grants (IIS-0713142 and IIS-1016614).

References

[1] A. Acharya, A. Rawal, R. J. Mooney, and E. R. Hruschka, *Using both supervised and latent shared topics for multitask learning*, in ECML PKDD, Part II, LNAI 8189, 2013, pp. 369–384.

[2] D. M. Blei and J. D. Mcauliffe, *Supervised topic models*, in Proc. of NIPS, 2007.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet Allocation*, JMLR, 3 (2003), pp. 993–1022.

[4] A. Bordes, L. Bottou, P. Gallinari, and J. Weston, *Solving multiclass support vector machines with larank*, in Proc. of ICML, 2007, pp. 89–96.

[5] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, *Fast kernel classifiers with online and active learning*, JMLR, 6 (2005), pp. 1579–1619.

[6] R. Caruana, *Multitask learning*, Machine Learning, 28 (1997), pp. 41–75.

[7] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, *Worst-case analysis of selective sampling for linear classification*, JMLR, 7 (2006), pp. 1205–1230.

[8] Y. S. Chan and H. T. Ng, *Domain adaptation with active learning for word sense disambiguation*, in Proc. of ACL, 2007, pp. 49–56.

[9] J. Chang and D. Blei, *Relational topic models for document networks*, in Proc. of AISTATS, 2009.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the*

EM algorithm, J. Royal Statistical Society. Series B (Methodological), 39 (1977), pp. 1–38.

[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, *ImageNet: A large-scale hierarchical image database*, in Proc. of CVPR, 2009, pp. 248–255.

[12] J. Donahue and K. Grauman, *Annotator rationales for visual recognition*, in Proc. of ICCV, 2011, pp. 1395–1402.

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, *Describing objects by their attributes*, in Proc. of CVPR, 2009, pp. 1778–1785.

[14] V. Ferrari and A. Zisserman, *Learning visual attributes*, in Proc. of NIPS, 2007.

[15] A. Harpale and Y. Yang, *Active learning for multi-task adaptive filtering*, in Proc. of ICML, Omnipress, 2010, pp. 431–438.

[16] P. Jain and A. Kapoor, *Active learning for large multi-class problems*, in Proc. of CVPR, 2009, pp. 762–769.

[17] R. Jenatton, J. Audibert, and F. Bach, *Structured variable selection with sparsity-inducing norms*, JMLR, 12 (2011), pp. 2777–2824.

[18] A.J. Joshi, F. Porikli, and N. Papanikolopoulos, *Multi-class active learning for image classification*, in Proc. of CVPR, 2009, pp. 2372–2379.

[19] G. Jun and J. Ghosh, *An efficient active learning algorithm with knowledge transfer for hyperspectral remote sensing data*, in Proc. of International Geosci. and Sens. Symposium, vol. 1, 2008, pp. I–52–I–55.

[20] S. Kim and E. P. Xing, *Tree-guided group lasso for multi-task regression with structured sparsity*, in Proc. of ICML, 2010, pp. 543–550.

[21] A. Kovashka, S. Vijayanarasimhan, and K. Grauman, *Actively selecting annotations among objects and attributes*, in Proc. of ICCV, 2011, pp. 1403–1410.

[22] C. H. Lampert, H. Nickisch, and S. Harmeling, *Learning to detect unseen object classes by betweenclass attribute transfer*, in Proc. of CVPR, 2009, pp. 951–958.

[23] Y. Low, D. Agarwal, and A. J. Smola, *Multiple domain user personalization*, in Proc. of KDD, 2011, pp. 123–131.

[24] D. G. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60 (2004), pp. 91–110.

[25] R. M. Neal and G. E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, 1999.

[26] S. J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Transactions on Knowledge and Data Engineering, 22 (2010), pp. 1345–1359.

[27] A. Passos, P. Rai, J. Wainer, and H. Daumé III, *Flexible modeling of latent task structures in multitask learning*, in Proc. of ICML, 2012, pp. 1103–1110.

[28] G.J. Qi, Xian-Sheng H., Yong R., Jinhui T., and Hong-Jiang Z., *Two-dimensional active learning for image classification*, in Proc. of CVPR, 2008, pp. 1–8.

[29] A. Quattoni, S. Wang, L. P Morency, M. Collins, and T. Darrell, *Hidden-state conditional random fields*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.

[30] P. Rai, A. Saha, H. Daumé, III, and S. Venkatasubramanian, *Domain adaptation meets active learning*, in Proc. of NAACL HLT Workshop on Active Learning for Natural Language Processing, 2010, pp. 27–32.

[31] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, *Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora*, in Proc. of EMNLP, 2009, pp. 248–256.

[32] S. Robertson and I. Soboroff, *The TREC 2002 filtering track report*, in Text Retrieval Conference, 2002.

[33] N. Roy and A. K. McCallum, *Toward optimal active learning through sampling estimation of error reduction*, in Proc. of ICML, 2001, pp. 441–448.

[34] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, *Statistical topic models for multi-label document classification*, CoRR, abs/1107.2462, 2011.

[35] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, *Labelme: A database and web-based tool for image annotation*, 2008.

[36] A. Saha, P. Rai, H. Daum III, and S. Venkatasubramanian, *Online learning of multiple tasks and their relationships.*, JMLR - Proceedings Track, 15 (2011), pp. 643–651.

[37] B. Settles, *Active learning literature survey*, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[38] X. Shi, W. Fan, and J. Ren, *Actively transfer domain knowledge*, in Proc. of ECML PKDD - Part II, 2008, pp. 342–357.

[39] S. Vijayanarasimhan, P. Jain, and K. Grauman, *Hashing hyperplane queries to near points with applications to large-scale active learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 36 (2014), pp. 276–288.

[40] C. Wang, J. W. Paisley, and D. M. Blei, *Online variational inference for the hierarchical Dirichlet process*, JMLR - Proceedings Track, 15 (2011), pp. 752–760.

[41] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, *Feature hashing for large scale multitask learning*, in Proc. of ICML, 2009, pp. 1113–1120.

[42] O. F. Zaidan, J. Eisner, and C. Piatko, *Machine learning with annotator rationales to reduce annotation cost*, in Proc. of the NIPS Workshop on Cost Sensitive Learning, 2008.

[43] J. Zhang, Z. Ghahramani, and Y. Yang, *Flexible latent variable models for multi-task learning*, Machine Learning, 73 (2008), pp. 221–242.

[44] J. Zhu, A. Ahmed, and E. P. Xing, *MedLDA: maximum margin supervised topic models for regression and classification*, in Proc. of ICML, 2009, pp. 1257–1264.