Proceedings, The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)

Learning Supervised Topic Models from Crowds

Filipe Rodrigues
University of Coimbra
fmpr@dei.uc.pt

Bernardete Ribeiro
University of Coimbra
bribeiro@dei.uc.pt

Mariana Lourenço
University of Coimbra
mrlouren@student.dei.uc.pt

Francisco Pereira
Massachusetts Institute of Technology
camara@mit.edu

Abstract

The growing need to analyze large collections of documents has led to great developments in topic modeling. Since documents are frequently associated with other related variables, such as labels or ratings, much interest has been placed on supervised topic models. However, the nature of most annotation tasks, prone to ambiguity and noise, often with high volumes of documents, deem learning under a single-annotator assumption unrealistic or unpractical for most real-world applications. In this paper, we propose a supervised topic model that accounts for the heterogeneity and biases among different annotators that are encountered in practice when learning from crowds. We develop an efficient stochastic variational inference algorithm that is able to scale to very large datasets, and we empirically demonstrate the advantages of the proposed model over state of the art approaches.

Introduction

Topic models, such as latent Dirichlet allocation (LDA), allow us to analyze large collections of documents, by revealing their underlying themes, or topics, and how each document exhibits them (Blei, Ng, and Jordan 2003). Therefore, it is not surprising that topic models have become a standard tool in machine learning, with many applications that transcend their original purpose of modeling textual data, such as analyzing images (Fei-Fei and Perona 2005; Wang, Blei, and Fei-Fei 2009), videos (Niebles, Wang, and Fei-Fei 2008), survey data (Erosheva, Fienberg, and Joutard 2007) or social networks data (Airoldi et al. 2007).

Since documents are frequently associated with other variables such as labels, tags or ratings, much interest has been placed on supervised topic models (Mcauliffe and Blei 2008), which allow the use of that extra information to "guide" the topics discovery. By jointly learning the topics distributions and a regression or classification model, supervised topic models have been shown to outperform the separate use of their unsupervised analogues with an external regression/classification algorithm (Wang, Blei, and Fei-Fei 2009; Zhu, Ahmed, and Xing 2012).

Supervised topics models are then state-of-the-art approaches for predicting target variables associated with complex high-dimensional data, such as documents or images. Unfortunately, the size of modern datasets make the use of a single annotator unrealistic and unpractical for the majority of the real-world applications that involve some form of human labeling. For instance, the popular Reuters-21578 benchmark dataset was categorized by a group of personnel from Reuters Ltd and Carnegie Group, Inc. Similarly, the LabelMe[1] project asks volunteers to annotate images from a large collection using an online tool. Hence, it is seldom the case where a single oracle labels an entire collection.

Furthermore, the Web, through its social nature, also exploits the wisdom of crowds to annotate large collections of documents and images. By categorizing texts, tagging images or rating products, Web users are generating large volumes of labeled content. However, when learning supervised models from crowds, the quality of labels can vary a lot due to task subjectivity and differences in annotator reliability (or bias) (Snow et al. 2008; Rodrigues, Pereira, and Ribeiro 2013). It is therefore essential to account for these issues when learning from this increasingly common type of data. Hence, the interest of researchers on building models that take the reliabilities of different annotators into consideration and mitigate the effect of their biases has spiked during the last few years (e.g. (Welinder et al. 2010; Yan et al. 2014)).

The increasing popularity of crowdsourcing platforms like Amazon Mechanical Turk (AMT) has further contributed to the recent developments in learning from crowds. This kind of platforms offer a fast, scalable and inexpensive solution for labeling large amounts of data. However, their heterogeneous nature in terms of contributors makes their straightforward application prone to many sorts of labeling noise and bias. Hence, a careless use of crowdsourced data as training data risks generating flawed models.

In this paper we propose a fully generative supervised

[1] http://labelme.csail.mit.edu

topic model that is able to account for the different reliabilities of multiple annotators and correct their biases. The proposed model is then capable of jointly modeling the words in documents as arising from a mixture of topics, the latent true labels as a result of the empirical distribution over topics of the documents, and the labels of the multiple annotators as noisy versions of that latent ground truth. Although we focus on multi-class classification problems, the same rationale can be applied to regression problems. Since the majority of the tasks for which multiple annotators are used generally involve complex data such as text, images and video, by developing a multi-annotator supervised topic model we are contributing with a powerful tool for learning predictive models of complex high-dimensional data from crowds.

Given that the increasing sizes of modern datasets can pose a problem for obtaining human labels as well as for Bayesian inference, we propose an efficient stochastic variational inference algorithm (Hoffman et al. 2013) that is able to scale to very large datasets. We empirically show, using both simulated and real multiple-annotator labels obtained from AMT for popular text and image collections, that the proposed model is able to outperform other state-of-the-art approaches. We further show the computational and predictive advantages of the stochastic variational inference algorithm over its batch counterpart.

State of the art

Latent Dirichlet allocation (LDA) soon proved to be a powerful tool for modeling documents (Blei, Ng, and Jordan 2003) and images (Fei-Fei and Perona 2005), by extracting their underlying topics. However, the need to model the relationship between documents and labels quickly gave rise to many supervised variants of LDA. One of the first notable works was that of (Mcauliffe and Blei 2008) in developing supervised LDA (sLDA). By extending LDA through the inclusion of a response variable that is linearly dependent on the mean topic-assignments of the words in a document, sLDA is able to jointly model the documents and their responses, in order to find latent topics that will best predict the response variables for future unlabeled documents. Although initially developed for general continuous response variables, (Wang, Blei, and Fei-Fei 2009) later extended sLDA to classification problems, by modeling the relationship between topic-assignments and labels with a softmax function.

There are several ways in which document classes can be included in LDA. The most natural one in this setting is probably the sLDA approach, since the classes are directly dependent on the empirical topic mixture distributions. This approach is coherent with the generative perspective of LDA but, nevertheless, several discriminative alternatives also exist. For example, DiscLDA (Lacoste-Julien, Sha, and Jordan 2009) introduces a class-dependent linear transformation on the topic mixture proportions, whose parameters are estimated by maximizing the conditional likelihood of response variables. (Ramage et al. 2009) propose Labeled-LDA, a variant of LDA that incorporates supervision by constraining the topic model to use only the topics that correspond to a document's label set. While this has the advantage of allowing multiple labels per document, it is restrictive in the sense that the number of topics needs to be the same as the number of possible labels.

The approaches discussed so far rely on likelihood-based estimation procedures. The work of (Zhu, Ahmed, and Xing 2012) contrasts with these approaches by proposing MedLDA, a supervised topic model that utilizes the max-margin principle for estimation. Despite its margin-based advantages, MedLDA looses the probabilistic interpretation of the document classes given the topic mixture distributions. On the contrary, this paper proposes a fully generative probabilistic model of the labels of multiple annotators and the words in the documents.

Learning from multiple annotators is an increasingly important research topic. Since the early work of (Dawid and Skene 1979), who attempted to obtain point estimates of the error rates of patients given repeated but conflicting responses to various medical questions, many approaches have been proposed. These usually rely on latent variable models. For example, (Smyth et al. 1995) proposed a model to estimate the ground truth from the labels of multiple experts, which is then used to train a classifier.

While earlier works usually focused on estimating the ground truth and the error rates of different annotators, recent works are more focused on the problem of learning a classifier. This idea was explored in (Raykar et al. 2010), who proposed an approach for jointly learning the levels of expertise of different annotators and the parameters of a logistic regression classifier, by modeling the ground truth labels as latent variables. This work was later extended by (Yan et al. 2014) by considering the dependencies of the annotators' labels on the instances they are labeling, and also by (Rodrigues, Pereira, and Ribeiro 2014) through the use of Gaussian process classifiers. The model proposed in this paper shares the same intuition with this line of work, and models the true labels as latent variables. However, it differs significantly by using a fully Bayesian approach for estimating the reliabilities and biases of the different annotators. Furthermore, it considers the problems of learning a low-dimensional representation of the input data (through topic modeling) and modeling the answers of multiple annotators jointly, providing an efficient stochastic variational inference algorithm.

Approach

In this section we develop a multi-class supervised topic model with multiple annotators. We start by deriving a (*batch*) variational inference algorithm for approximating the posterior distribution over the latent variables and an algorithm to estimate the model parameters. We then develop a stochastic variational inference algorithm that gives the model the capability of handling large collections of documents. Finally, we show how to use the learned model to classify new documents.

Proposed model

Let $\mathcal{D} = \{\mathbf{w}^d, \mathbf{y}^d\}_{d=1}^D$ be an annotated corpus of size D, where each document $\mathbf{w}^d$ is given a set of labels $\mathbf{y}^d =$

$\{y_r^d\}_{r=1}^{R_d}$ from R_d distinct annotators. We can take advantage of the inherent topical structure of documents and model their words as arising from a mixture of topics, each being defined as a distribution over the words in a vocabulary, as in LDA. In LDA, the n^{th} word, w_n^d, in a document d is provided a discrete topic-assignment z_n^d, which is drawn from the documents' distribution over topics θ^d. This allows us to build lower-dimensional representations of documents, which we can explore to build classification models by assigning coefficients $\boldsymbol{\eta}$ to the mean topic-assignment of the words in the document, $\bar{z}^d$, and applying a softmax function in order to obtain a distribution over classes.

Unfortunately, a direct mapping between document classes and the labels provided by the different annotators in a multiple-annotator setting would correspond to assuming that they are all equally reliable, an assumption that is violated in practice, as previous works clearly demonstrate (e.g. (Snow et al. 2008; Rodrigues, Pereira, and Ribeiro 2013)). Hence, we assume the existence of a latent ground truth class, and model the labels from the different annotators using a noise model that states that, given a true class c, each annotator r provides the label l with some probability $\pi_{c,l}^r$. Hence, by modeling $\boldsymbol{\pi}^r$ we are in fact modeling a per-annotator confusion matrix, which allows us to account for their different levels of expertise and correct their potential biases.

The generative process of the proposed model can then be summarized as follows:

1. For each annotator r
   (a) For each class c
      i. Draw reliability parameter $\pi_c^r|\omega \sim Dir(\omega)$
2. For each topic k
   (a) Draw topic distribution $\beta_k|\tau \sim Dir(\tau)$
3. For each document d
   (a) Draw topic proportions $\theta^d|\alpha \sim Dir(\alpha)$
   (b) For the n^{th} word
      i. Draw topic assignment $z_n^d|\theta^d \sim Mult(\theta^d)$
      ii. Draw word $w_n^d|z_n^d, \boldsymbol{\beta} \sim Mult(\beta_{z_n^d})$
   (c) Draw latent (true) class $c^d|\mathbf{z}^d, \boldsymbol{\eta} \sim Softmax(\bar{z}^d, \boldsymbol{\eta})$
   (d) For each annotator $r \in R_d$
      i. Draw annotator's label $y^{d,r}|c^d, \boldsymbol{\pi}^r \sim Mult(\pi_{c^d}^r)$

where R_d denotes the set of annotators that labeled the d^{th} document, $\bar{z}^d = \frac{1}{N_d}\sum_{n=1}^{N_d} z_n^d$, and the softmax is given by:

$$p(c^d|\bar{z}^d, \boldsymbol{\eta}) = \frac{\exp(\eta_c^T \bar{z}^d)}{\sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d)}.$$

Figure 1 shows a graphical model representation of the proposed model, where K denotes the number of topics, C is the number of classes, R is the total number of annotators and N_d is the number of words in the document d. Notice that we included a Dirichlet prior over the topics β_k to produce a smooth posterior and control sparsity. Similarly, instead of computing maximum likelihood or MAP estimates for the annotators reliability parameters π_c^r, we place a Dirichlet prior over these variables and perform (approximate) Bayesian inference. This contrasts with previous works on learning from crowds (Raykar et al. 2010; Yan et al. 2010).

Variational param.	Original param.
ξ_c^r	π_c^r
ζ_k	β_k
γ^d	θ^d
λ^d	c^d
ϕ_n^d	z_n^d

Table 1: Correspondence between variational parameters and the original parameters.

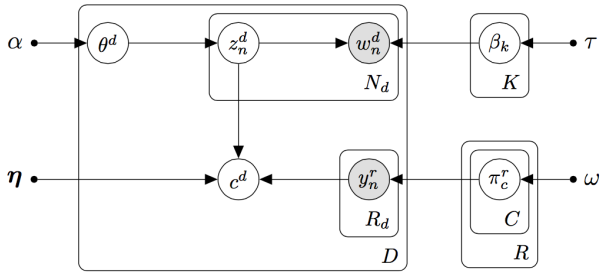

Figure 1: Graphical model representation of the proposed model.

Approximate inference

Given a dataset $\mathcal{D}$, the goal of inference is to compute the posterior distribution of the per-document topic proportions θ^d, the per-word topic assignments z_n^d, the per-topic distribution over words β_k, the per-document latent true class c^d, and the per-annotator confusion parameters $\boldsymbol{\pi}^r$. As with LDA, computing the exact posterior distribution of the latent variables is computationally intractable. Hence, we employ mean-field variational inference to perform approximate Bayesian inference.

Variational inference methods seek to minimize the KL divergence between the variational and the true posterior distribution. We assume a fully-factorized (mean-field) variational distribution of the form:

$$q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}) = \left(\prod_{r=1}^{R}\prod_{c=1}^{C} q(\pi_c^r|\xi_c^r)\right) \times \left(\prod_{i=1}^{K} q(\beta_i|\zeta_i)\right)\prod_{d=1}^{D} q(\theta^d|\gamma^d)q(c^d|\lambda^d)\prod_{n=1}^{N_d} q(z_n^d|\phi_n^d),$$

where $\boldsymbol{\xi}_{1:R}$, $\boldsymbol{\zeta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}_{1:D}$ are variational parameters. Table 1 shows the correspondence between variational parameters and the original parameters.

Let $\Theta = \{\alpha, \tau, \omega, \boldsymbol{\eta}\}$ denote the model parameters. Following (Jordan et al. 1999), the KL minimization can be equivalently formulated as maximizing the following lower

bound on the log marginal likelihood,

$$\begin{aligned}
&\log p(\mathbf{w}_{1:D}, \mathbf{y}_{1:D}|\Theta) \\
&\quad = \log \int \sum_{\mathbf{z},\boldsymbol{c}} q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}) \\
&\quad \times \frac{p(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \mathbf{w}_{1:D}, \mathbf{y}_{1:D}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}|\Theta)}{q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R})} \, d\boldsymbol{\theta} \, d\boldsymbol{\beta} \, d\boldsymbol{\pi}_{1:R} \\
&\quad \geqslant \mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \mathbf{w}_{1:D}, \mathbf{y}_{1:D}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}|\Theta)] \\
&\quad + \mathbb{E}_q[\log q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R})] \\
&\quad = \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}_{1:D}, \boldsymbol{\lambda}, \boldsymbol{\zeta}, \boldsymbol{\xi}_{1:R}|\Theta), \qquad (1)
\end{aligned}$$

which we maximize using coordinate ascent.

Optimizing $\mathcal{L}$ w.r.t. γ and ζ gives the same coordinate ascent updates as in (Blei, Ng, and Jordan 2003):

$$\gamma_i^d = \alpha + \sum_{n=1}^{N_d} \phi_{n,i}^d \qquad (2)$$

$$\zeta_{i,j} = \tau + \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_{n,j}^d \phi_{n,i}^d. \qquad (3)$$

The variational Dirichlet parameters ξ can be optimized by collecting only the terms in $\mathcal{L}$ that contain ξ:

$$\begin{aligned}
\mathcal{L}_{[\xi]} = &\sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{l=1}^{C} \mathbb{E}_q[\log \pi_{c,l}^r] \Big(\omega + \sum_{d=1}^{D_r} \lambda_c^d y_l^{d,r} - \xi_{c,l}^r \Big) \\
&- \sum_{r=1}^{R} \sum_{c=1}^{C} \log \Gamma \Big(\sum_{t=1}^{C} \xi_{c,t}^r \Big) + \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{l=1}^{C} \log \Gamma(\xi_{c,l}^r),
\end{aligned}$$

where D_r denotes the documents labeled by the r^{th} annotator, $\mathbb{E}_q[\log \pi_{c,l}^r] = \Psi(\xi_{c,l}^r) - \Psi(\sum_{t=1}^{C} \xi_{c,t}^r)$, and $\Gamma(\cdot)$ and $\Psi(\cdot)$ are the gamma and digamma functions, respectively. Taking derivatives of $\mathcal{L}_{[\xi]}$ w.r.t. ξ and setting them to zero, yields the following update:

$$\xi_{c,l}^r = \omega + \sum_{d=1}^{D_r} \lambda_c^d y_l^{d,r}. \qquad (4)$$

Similarly, the coordinate ascent updates for the documents distribution over classes λ can be found by considering the terms in $\mathcal{L}$ that contain λ:

$$\begin{aligned}
\mathcal{L}_{[\lambda]} = &\sum_{d=1}^{D} \sum_{l=1}^{C} \lambda_l^d \eta_l^T \bar{\phi}^d - \sum_{l=1}^{C} \lambda_l^d \log \lambda_l^d \\
&+ \sum_{d=1}^{D} \sum_{r=1}^{R_d} \sum_{l=1}^{C} \sum_{c=1}^{C} \lambda_l^d \, y_c^{d,r} \, \mathbb{E}_q[\log \pi_{l,c}^r],
\end{aligned}$$

where $\bar{\phi}^d = \frac{1}{N_d} \sum_{n=1}^{N_d} \phi_n^d$. Adding the necessary Lagrange multipliers to ensure that $\sum_{l=1}^{C} \lambda_l^d = 1$ and setting the derivatives w.r.t. λ_l^d to zero gives the following update:

$$\lambda_l^d \propto \exp\Big(\eta_l^T \bar{\phi}^d + \sum_{r=1}^{R_d} \sum_{c=1}^{C} y_c^{d,r} \, \mathbb{E}_q[\log \pi_{l,c}^r] \Big). \qquad (5)$$

Observe how the variational distribution over the true classes results from a combination between the dot product of the inferred mean topic assignment $\bar{\phi}^d$ with the coefficients $\boldsymbol{\eta}$ and the labels $\mathbf{y}$ from the multiple annotators "weighted" by their expected log probability $\mathbb{E}_q[\log \pi_{l,c}^r]$.

The main difficulty of applying standard variational inference methods to the proposed model is the non-conjugacy between the distribution of the mean topic-assignment $\bar{z}^d$ and the softmax. Namely, in the expectation

$$\mathbb{E}_q[\log p(c^d|\bar{z}^d, \boldsymbol{\eta})] = \mathbb{E}_q[\eta_{c^d}^T \bar{z}^d] - \mathbb{E}_q\Big[\log \sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d) \Big],$$

the second term is intractable to compute. We can make progress by applying Jensen's inequality to bound it as follows:

$$\begin{aligned}
-\mathbb{E}_q\Big[\log \sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d) \Big] &\geqslant -\log \sum_{l=1}^{C} \mathbb{E}_q[\exp(\eta_l^T \bar{z}^d)] \\
&= -\log \sum_{l=1}^{C} \prod_{j=1}^{N_d} \left(\phi_j^d\right)^T \exp\Big(\eta_l \frac{1}{N_d}\Big) \\
&= -\log(a^T \phi_n^d),
\end{aligned}$$

where $a \triangleq \sum_{l=1}^{C} \exp(\frac{\eta_l}{N_d}) \prod_{j=1, j\neq n}^{N_d} \left(\phi_j^d\right)^T \exp\Big(\frac{\eta_l}{N_d}\Big)$, which is constant w.r.t. ϕ_n^d. This local variational bound can be made tight by noticing that $\log(x) \leqslant \epsilon^{-1} x + \log(\epsilon) - 1, \forall x > 0, \epsilon > 0$, where equality holds if and only if $x = \epsilon$. Hence, given the current parameter estimates $(\phi_n^d)^{old}$, if we set $x = a^T \phi_n^d$ and $\epsilon = a^T (\phi_n^d)^{old}$ then, for an individual parameter ϕ_n^d, we have that:

$$\begin{aligned}
&-\mathbb{E}_q\Big[\log \sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d) \Big] \\
&\qquad \geqslant -(a^T (\phi_n^d)^{old})^{-1} (a^T \phi_n^d) - \log(a^T (\phi_n^d)^{old}) + 1.
\end{aligned}$$

Using this local bound to approximate the expectation of the log-sum-exp term, and taking derivatives of the evidence lower bound w.r.t. ϕ_n with the constraint that $\sum_{i=1}^{K} \phi_{n,i}^d = 1$, yields the following fix-point update:

$$\begin{aligned}
\phi_{n,i}^d \propto \exp\Bigg(&\Psi(\gamma_i) + \sum_{j=1}^{V} w_{n,j}^d \Big(\Psi(\zeta_{i,j}) - \Psi\Big(\sum_{k=1}^{V} \zeta_{i,k} \Big) \Big) \\
&+ \frac{\sum_{l=1}^{C} \lambda_l^d \eta_{l,i}}{N_d} - (a^T (\phi_n^d)^{old})^{-1} a_i \Bigg). \qquad (6)
\end{aligned}$$

where V denotes the size of the vocabulary. Notice how the per-word variational distribution over topics ϕ depends on the variational distribution over the true class label λ.

The variational inference algorithm iterates between equations 2-6 until the evidence lower bound, eq. 1, converges. The supplementary material provides additional details on the derivation of this algorithm[2].

[2]Supplementary material available at: http://amilab.dei.uc.pt/fmpr/maslda-supp-mat.pdf

Parameter estimation

The model parameters are $\Theta = \{\alpha, \tau, \omega, \boldsymbol{\eta}\}$. For the sake of simplicity we assume the parameters α, τ and ω of the Dirichlet priors to be fixed, and only estimate the coefficients $\boldsymbol{\eta}$ using a variational EM algorithm. Therefore, in the E-step we use the variational inference algorithm from approximate inference section to estimate the posterior distribution of the latent variables, and in the M-step we find maximum likelihood estimates of $\boldsymbol{\eta}$ by maximizing the evidence lower bound $\mathcal{L}$. Unfortunately, taking derivatives of $\mathcal{L}$ w.r.t. $\boldsymbol{\eta}$ does not yield a closed-form solution, hence we use a numerical method, namely L-BFGS (Nocedal and Wright 2006), to find an optimum. The objective function and gradients are given by

$$\mathcal{L}_{[\eta]} = \sum_{d=1}^{D} \left(\sum_{l=1}^{C} \lambda_l^d \eta_l^T \bar{\phi}^d - \log \sum_{l=1}^{C} b_l^d \right)$$

$$\nabla_{\eta_{l,i}} = \sum_{d=1}^{D} \left(\lambda_{l,i}^d \bar{\phi}_i^d - \frac{b_l^d}{\sum_{t=1}^{C} b_t^d} \times \sum_{n=1}^{N_d} \frac{\frac{1}{N_d} \phi_{n,i}^d \exp(\frac{1}{N_d} \eta_{l,i})}{\sum_{j=1}^{K} \phi_{n,j}^d \exp(\frac{1}{N_d} \eta_{l,j})} \right),$$

where, for convenience, we defined the following variable: $b_l^d \triangleq \prod_{n=1}^{N_d} \left(\sum_{i=1}^{K} \phi_{n,i}^d \exp\left(\frac{1}{N_d} \eta_{l,i}\right)\right)$.

Stochastic variational inference

In the "approximate inference" section, we proposed a batch coordinate ascent algorithm for doing variational inference in the proposed model. This algorithm iterates between analyzing every document in the corpus to infer the local hidden structure, and estimating the global hidden variables. However, this can be inefficient for large datasets, since it requires a full pass through the data at each iteration before updating the global variables. In this section we develop a stochastic variational inference algorithm (Hoffman et al. 2013), which follows noisy estimates of the gradients of the evidence lower bound $\mathcal{L}$.

Based on the theory of stochastic optimization (Robbins and Monro 1951), we can find unbiased estimates of the gradients by subsampling a document (or a mini-batch of documents) from the corpus, and using it to compute the gradients as if that document was observed D times. Hence, given an uniformly sampled document d, we use the current posterior distributions of the global latent variables, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_{1:R}$, and the current coefficient estimates $\boldsymbol{\eta}$, to compute the posterior distribution over the local hidden variables θ^d, $\mathbf{z}^d$ and c^d using eqs. 2, 6 and 5 respectively. These posteriors are then used to update the global variational parameters, $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}_{1:R}$ by taking a step of size ρ_t in the direction of the noisy estimates of the natural gradients.

Algorithm 1 describes a stochastic variational inference algorithm for the proposed model. Given an appropriate schedule for the learning rates $\{\rho_t\}$, such that $\sum_t \rho_t$ and $\sum_t \rho_t^2 < \infty$, the stochastic optimization algorithm is guaranteed to converge to a local maximum of the evidence lower bound (Robbins and Monro 1951).

Algorithm 1 Stochastic variational inference

1: Initialize $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\phi}_{1:D}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\boldsymbol{\zeta}^{(0)}$, $\boldsymbol{\xi}_{1:R}^{(0)}$, $t = 0$
2: **repeat**
3: Set t = t + 1.
4: Sample a document $\mathbf{w}^d$ uniformly from the corpus.
5: **repeat**
6: Compute ϕ_n^d using eq. 6, for $n \in \{1..N_d\}$.
7: Compute γ^d using eq. 2.
8: Compute λ^d using eq. 5.
9: **until** local parameters ϕ_n^d, γ^d and λ^d converge.
10: Compute step-size $\rho_t = (t + delay)^{-\kappa}$.
11: Update topics variational parameters

$$\zeta_{i,j}{}^{(t)} = (1 - \rho_t)\zeta_{i,j}^{(t-1)} + \rho_t \left(\tau + D \sum_{n=1}^{N_d} w_{n,j}^d \phi_{n,i}^d \right).$$

12: Update annotators confusion parameters

$$\xi_{c,l}^r{}^{(t)} = (1 - \rho_t)\xi_{c,l}^r{}^{(t-1)} + \rho_t \left(\omega + D\, \lambda_c^d\, y_l^{d,r}\right).$$

13: **until** global convergence criterion is met.

Document classification

In order to make predictions for a new (unlabeled) document d, we start by computing the approximate posterior distribution over the latent variables θ^d and $\mathbf{z}^d$. This can be achieved by dropping the terms that involve y, c and π from the model's joint distribution (since, at prediction time, the multi-annotator labels are no longer observed) and averaging over the estimated topics distributions. Letting the topics distribution over words inferred during training be $q(\boldsymbol{\beta}|\boldsymbol{\zeta})$, the joint distribution for a single document is now simply given by

$$p(\theta^d, \mathbf{z}^d) = \int q(\boldsymbol{\beta}|\boldsymbol{\zeta}) p(\theta^d|\alpha) \prod_{n=1}^{N_d} p(z_n^d|\theta^d) p(w_n^d|z_n^d, \boldsymbol{\beta}) d\boldsymbol{\beta}.$$

Deriving a mean-field variational inference algorithm for computing the posterior over $q(\theta^d, \mathbf{z}^d) = q(\theta^d|\gamma^d) \prod_{n=1}^{N_d} q(z_n^d|\phi_n^d)$ results in the same fixed-point updates as in LDA (Blei, Ng, and Jordan 2003) for γ_i^d and $\phi_{n,i}^d$. Using the inferred posteriors and the coefficients $\boldsymbol{\eta}$ estimated during training, we can make predictions as follows:

$$c_*^d = \arg\max_c \eta_c^T \bar{\phi}^d. \tag{7}$$

This is equivalent to making predictions in sLDA (Wang, Blei, and Fei-Fei 2009).

Experiments

In this section, the proposed model, multi-annotator supervised LDA (MA-sLDA), is validated using both simulated annotators on popular corpora and using real multiple-annotator labels obtained from Amazon Mechanical Turk.[3]

[3] Source code and datasets used are available at: http://amilab.dei.uc.pt/fmpr/

Simulated annotators

In order to first validate the proposed model in a slightly more controlled environment, the well-known 20-Newsgroups benchmark corpus (Lang 1995) was used by simulating multiple annotators with different levels of expertise. The 20-Newsgroups consists of twenty thousand messages taken from twenty newsgroups, and is divided in six super-classes, which are, in turn, partitioned in several subclasses. For this first set of experiments, only the four most populated super-classes were used: "computers", "science", "politics" and "recreative". The preprocessing of the documents consisted of stemming and stop-words removal. After that, 75% of the documents were randomly selected for training and the remaining 25% for testing.

The different annotators were simulated by sampling their answers from a multinomial distribution, where the parameters are given by the lines of the annotators' confusion matrices. Hence, for each annotator r, we start by pre-defining a confusion matrix $\boldsymbol{\pi}^r$ with elements $\pi^r_{c,l}$, which correspond to the probability that the annotators' answer is l given that the true label is c, $p(y^r_i = l|c)$. Then, the answers are sampled i.i.d. from $y^r_i \sim Mult(\pi^r_{c,l})$. This procedure was used to simulate 5 different annotators with the following accuracies: 0.737, 0.468, 0.284, 0.278, 0.260. The distributions of the accuracies of the different annotators among the different classes are relatively uniform. In this experiment, no repeated labelling was used. Hence, each annotator only labels roughly one-fifth of the data. When compared to the ground truth, the simulated answers revealed an accuracy of 0.405. See Table 2 for an overview of the details of the datasets used.

Both the *batch* and the stochastic variational inference (*svi*) versions of the proposed model (MA-sLDA) are compared with the following baselines:

- *LDA + LogReg (mv)*: This baseline corresponds to applying unsupervised LDA to the data, and learning a logistic regression classifier on the inferred topics distributions of the documents. The labels from the different annotators were aggregated using majority voting (mv). Notice that, when there is a single annotator label per instance, majority voting is equivalent to using that label for training. This is the case of the 20-Newsgroups' simulated annotators, but the same does not apply for the experiments with Amazon Mechanical Turk.
- *LDA + Raykar*: For this baseline, the model of (Raykar et al. 2010) was applied using the documents' topic distributions inferred by LDA as features.
- *LDA + Rodrigues*: This baseline is similar to the previous one, but uses the model of (Rodrigues, Pereira, and Ribeiro 2013) instead.
- *Blei 2003 (mv)*: The idea of this baseline is to replicate a popular state-of-the-art approach for document classification. Hence, the approach of (Blei, Ng, and Jordan 2003) was used. It consists of applying LDA to extract the documents' topics distributions, which are then used to train a SVM. Similarly to the previous approach, the labels from the different annotators were aggregated using majority voting (mv).

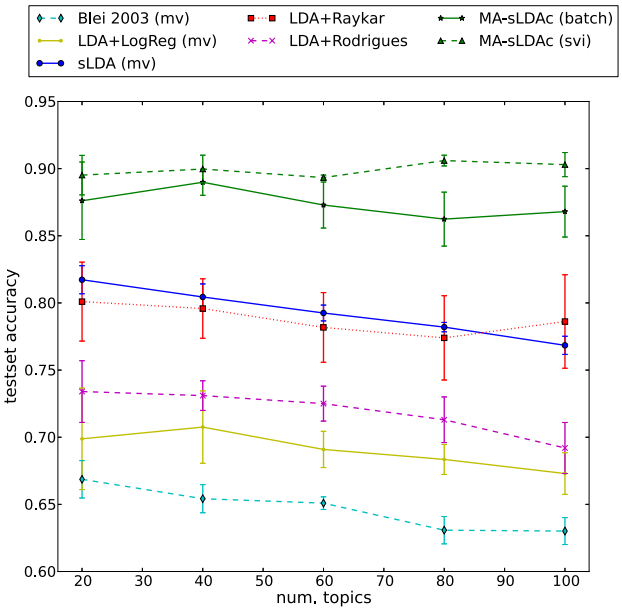

Figure 2: Average testset accuracy (over 5 runs; $\pm$ stddev.) of the different approaches on the 20-Newsgroups data.

- *sLDA (mv)*: This corresponds to using sLDA (Wang, Blei, and Fei-Fei 2009) with the labels obtained by performing majority voting (mv) on the annotators' answers.

For all the experiments the hyper-parameters α, τ and ω were set using a simple grid search in the collection $\{0.01, 0.1, 1.0, 10.0\}$. The same approach was used to optimize the hyper-parameters of the all the baselines. For the *svi* algorithm, different mini-batch sizes and forgetting rates κ were tested. For the 20-Newsgroup dataset, the best results were obtained with a mini-batch size of 500 and $\kappa = 0.6$. The $delay$ was kept at 1. The results are shown in Figure 2 for different numbers of topics, where we can see that the proposed model outperforms all the baselines, being the *svi* version the one that performs best.

In order to assess the computational advantages of the stochastic variational inference (*svi*) over the *batch* algorithm, the log marginal likelihood (or log evidence) was plotted against the number of iterations. Figure 3 shows this comparison. Not surprisingly, the *svi* version converges much faster to higher values of the log marginal likelihood when compared to the *batch* version, which reflects the efficiency of the *svi* algorithm.

Amazon Mechanical Turk

In order to validate the proposed model in a real crowdsourcing setting, Amazon Mechanical Turk (AMT) was used to obtain labels from multiple annotators for two popular datasets: Reuters-21578 (Lewis 1997) and LabelMe (Russell et al. 2008).

Reuters-21578 is a collection of manually categorized newswire stories with labels such as Acquisitions, Crude-oil, Earnings or Grain. For this experiment, only the documents

Dataset	Num. classes	Train/test sizes	Annotators source	Num. answers per instance ($\pm$ stddev.)	Mean annotators accuracy ($\pm$ stddev.)	Maj. vot. accuracy
20 Newsgroups	4	11536/3846	Simulated	1.000 ± 0.000	0.405 ± 0.182	0.405
Reuters-21578	8	1800/5216	Mech. Turk	3.007 ± 1.019	0.568 ± 0.262	0.710
LabelMe	8	1000/1688	Mech. Turk	2.547 ± 0.576	0.692 ± 0.181	0.769

Table 2: Overall statistics of the datasets used in the experiments.

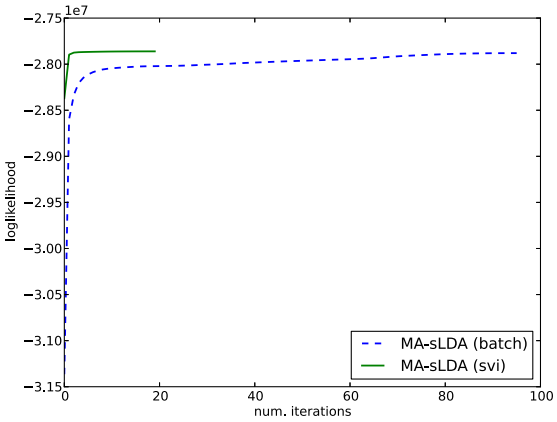

Figure 3: Comparison of the log marginal likelihood between the *batch* and the stochastic variational inference (*svi*) algorithms on the 20-Newsgroups corpus.

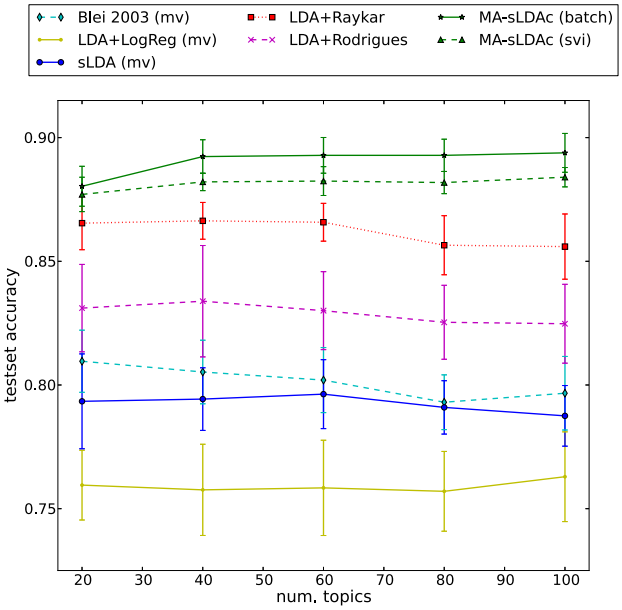

Figure 4: Average testset accuracy (over 30 runs; $\pm$ stddev.) of the different approaches on the Reuters data.

belonging to the widely-used ModApte split (Lewis et al. 2004) were considered with the additional constraint that the documents should have no more than one label. This resulted in a total of 7016 documents distributed among 8 classes. Of these, 1800 documents were submitted to AMT for multiple annotators to label, giving an average of 3.007 answers per document (see Table 2 for further details). The remaining 5216 documents were used for testing. The collected answers yield an average annotator accuracy of 56.8%. Applying majority voting to these answers reveals a ground truth accuracy of 71.0%.

The results obtained by the different approaches are given in Figure 4, where it can be seen that the proposed model (MA-sLDA) outperforms all the other approaches. For this dataset, the *svi* algorithm is using mini-batches of 300 documents.

The proposed model is also validated using a dataset from the computer vision domain: LabelMe (Russell et al. 2008). In contrast to the Reuters and Newsgroups corpora, LabelMe is an open online tool to annotate images. Hence, this experiment allows us to see how the proposed model generalises beyond non-textual data. Using the provided Matlab interface, we extracted a subset of the LabelMe data, consisting of all the 256 x 256 images with the categories: "highway", "inside city", "tall building", "street", "forest", "coast", "mountain" or "open country". This allowed us to collect a total of 2688 labeled images. Of these, 1000 images were given to AMT workers to classify with one of the classes above. Each image was labeled by an average of 2.547 workers, with a mean accuracy of 69.2%. When majority voting is applied to the collected answers, a ground truth accuracy of 71.0% is obtained.

The preprocessing of the images used is similar to the approach of (Fei-Fei and Perona 2005). It uses 128-dimensional SIFT (Lowe 1999) region descriptors selected by a sliding grid spaced at one pixel. This sliding grid extracts local regions of the image with sizes uniformly sampled between 16 x 16 and 32 x 32 pixels. The 128-dimensional SIFT descriptors produced by the sliding window are then fed to a k-means algorithm (with k=200) in order construct a vocabulary of 200 "visual words". This allows us to represent the images with a bag of visual words model.

With the purpose of comparing the proposed model with a popular state-of-the-art approach for image classification, for the LabelMe dataset, the following baseline was introduced:

- *Bosch 2006 (mv)*: This baseline is similar to one in (Bosch, Zisserman, and Muñoz 2006). The authors propose the use of pLSA to extract the latent topics, and the use of k-nearest neighbor (kNN) classifier using the documents' topics distributions. For this baseline, unsuper-

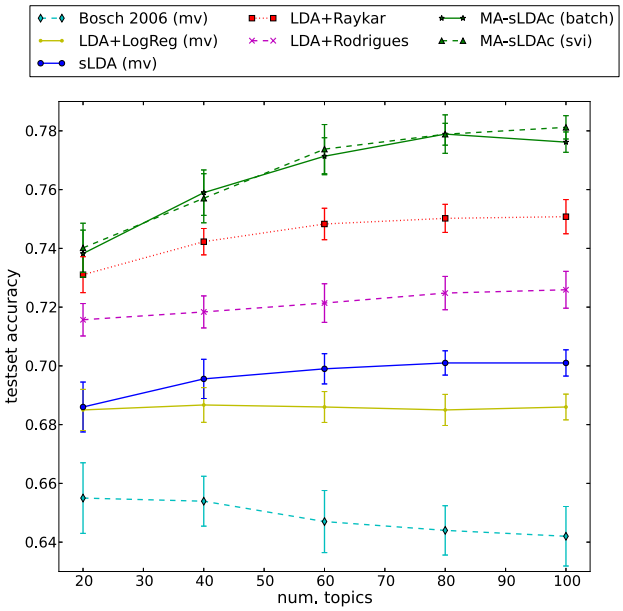

Figure 5: Average testset accuracy (over 30 runs; $\pm$ stddev.) of the different approaches on the LabelMe data.

vised LDA is used instead of pLSA, and the labels from the different annotators for kNN (with $k = 10$) are aggregated using majority voting (mv).

The results obtained by the different approaches for the LabelMe data are shown in Figure 5, where the *svi* version is using mini-batches of 200 documents.

Analyzing the results for the Reuters-21578 and LabelMe data, we can observe that the proposed model outperforms all the baselines, with slightly better accuracies for the *batch* version, especially in the Reuters data. Interestingly, the second best results are consistently obtained by the multi-annotator approaches, which highlights the need for accounting for the noise and biases of the answers of the different annotators.

Conclusion

This paper proposed a supervised topic model that is able to learn from multiple annotators and crowds, by accounting for their biases and different levels of expertise. Given the large sizes of modern datasets, and considering that the majority of the tasks for which crowdsourcing and multiple annotators are desirable candidates, generally involve complex high-dimensional data such as text and images, the proposed model constitutes a strong contribution for the multi-annotator paradigm. This model is then capable of jointly modeling the words in documents as arising from a mixture of topics, as well as the latent true labels and the (noisy) labels of the multiple annotators. We empirically showed, using simulated annotators on the 20-Newsgroups dataset and using real annotators from Amazon Mechanical Turk for Reuters-21578 and LabelMe data, that the proposed model is able to outperform state-of-the-art approaches. Finally, an efficient stochastic variational inference algorithm was described, which gives the proposed model the ability to scale to large datasets.

Given that the target variables associated with documents can be continuous, and also considering that documents can sometimes belong to more than one class, future work will explore the extension of the proposed model to regression and multi-label classification problems.

References

Airoldi, E.; Blei, D.; Fienberg, S.; and Xing, E. 2007. Combining stochastic block models and mixed membership for statistical network analysis. In *Statistical Network Analysis: Models, Issues, and New Directions*. Springer. 57–74.

Blei, D.; Ng, A.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Bosch, A.; Zisserman, A.; and Muñoz, X. 2006. Scene classification via plsa. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 517–530.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C* 28(1):20–28.

Erosheva, E.; Fienberg, S.; and Joutard, C. 2007. Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics* 1(2):346.

Fei-Fei, L., and Perona, P. 2005. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, 524–531. IEEE.

Hoffman, M.; Blei, D.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *J. Mach. Learn. Res.* 14(1):1303–1347.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T.; and Saul, L. 1999. An introduction to variational methods for graphical models. *Mach. Learn.* 37(2):183–233.

Lacoste-Julien, S.; Sha, F.; and Jordan, M. 2009. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, 897–904.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, 331–339.

Lewis, D.; Yang, Y.; Rose, T.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5:361–397.

Lewis, D. 1997. Reuters-21578 text categorization test collection, distribution 1.0.

Lowe, D. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, 1150–1157. Ieee.

Mcauliffe, J., and Blei, D. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

Niebles, J.; Wang, H.; and Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision* 79(3):299–318.

Nocedal, J., and Wright, S. 2006. *Numerical Optimization*. World Scientific.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–256. Association for Computational Linguistics.

Raykar, V.; Yu, S.; Zhao, L.; Valadez, G.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from Crowds. *Journal of Machine Learning Research* 1297–1322.

Robbins, H., and Monro, S. 1951. A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters* 1428–1436.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2014. Gaussian process classification and active learning with multiple annotators. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 433–441.

Russell, B.; Torralba, A.; Murphy, K.; and Freeman, W. 2008. Labelme: a database and web-based tool for image annotation. *International journal of computer vision* 77(1-3):157–173.

Smyth, P.; Fayyad, U.; Burl, M.; Perona, P.; and Baldi, P. 1995. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems*, 1085–1092.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and fast - but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 254–263.

Wang, C.; Blei, D.; and Fei-Fei, L. 2009. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1903–1910. IEEE.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, 2424–2432.

Yan, Y.; Rosales, R.; Fung, G.; Schmidt, M.; Valadez, G.; Bogoni, L.; Moy, L.; and Dy, J. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. *Journal of Machine Learning Research* 9:932–939.

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014. Learning from multiple annotators with varying expertise. *Mach. Learn.* 95(3):291–327.

Zhu, J.; Ahmed, A.; and Xing, E. 2012. Medlda: Maximum margin supervised topic models. *J. Mach. Learn. Res.* 13(1):2237–2278.