Early Gains Matter: A Case for Preferring Generative over Discriminative Crowdsourcing Models

Paul Felt, Eric Ringger, Kevin Seppi, Kevin Black, Robbie Haertel
Brigham Young University
Provo, UT 84602, USA
{paul_felt,kevin_black}@byu.edu, {ringger,kseppi}@cs.byu.edu, robbie.haertel@gmail.com

Abstract

In modern practice, labeling a dataset often involves aggregating annotator judgments obtained from crowdsourcing. State-of-the-art aggregation is performed via inference on probabilistic models, some of which are data-aware, meaning that they leverage features of the data (e.g., words in a document) in addition to annotator judgments. Previous work largely prefers discriminatively trained conditional models. This paper demonstrates that a data-aware crowdsourcing model incorporating a generative multinomial data model enjoys a strong competitive advantage over its discriminative log-linear counterpart in the typical crowdsourcing setting. That is, the generative approach is better except when the annotators are highly accurate in which case simple majority vote is often sufficient. Additionally, we present a novel mean-field variational inference algorithm for the generative model that significantly improves on the previously reported state-of-the-art for that model. We validate our conclusions on six text classification datasets with both human-generated and synthetic annotations.

1 Introduction

The success of supervised machine learning has created an urgent need for manually-labeled training datasets. Crowdsourcing allows human label judgments to be obtained rapidly and at relatively low cost. Micro-task markets such as Amazon's Mechanical Turk and CrowdFlower have popularized crowdsourcing by reducing the overhead required to distribute a job to a community of annotators (the "crowd"). However, crowdsourced judgments often suffer from high error rates. A common solution to this problem is to obtain multiple redundant human judgments, or annotations,[1] relying on the observation that, in aggregate, the ability of non-experts often rivals or exceeds that of experts by averaging over individual error patterns (Surowiecki, 2005; Snow et al., 2008; Jurgens, 2013).

For the purposes of this paper a *crowdsourcing model* is a model that infers, at a minimum, class labels y based on the evidence of one or more imperfect annotations a. A common baseline method aggregates annotations by *majority vote* but by so doing ignores important information. For example, some annotators are more reliable than others, and their judgments ought to be weighted accordingly. State-of-the-art crowdsourcing methods formulate probabilistic models that account for such side information and then apply standard inference techniques to the task of inferring ground truth labels from imperfect annotations.

Data-aware crowdsourcing models additionally account for the features x comprising each data instance (e.g., words in a document). The data can be modeled generatively by proposing a joint distribution $p(y,x,a)$. However, because of the challenge of accurately modeling complex data x, most previous work uses a discriminatively trained conditional model $p(y,a|x)$, hereafter referred to as a discriminative model. As Ng and Jordan (2001) explain, maximizing conditional log likelihood is a compu-

[1] We use the term *annotation* to identify human judgments and distinguish them from gold standard class *labels*.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 882–891,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

tationally convenient approximation to minimizing a discriminative 0-1 loss objective, giving rise to the common practice of referring to conditional models as discriminative.

Contributions. This paper challenges the popular preference for discriminative data models in the crowdsourcing literature by demonstrating that in typical crowdsourcing scenarios a generative model enjoys a strong advantage over its discriminative counterpart. We conduct, on both real and synthetic annotations, the first empirical comparison of structurally comparable generative and discriminative crowdsourcing models. The comparison is made fair by developing similar mean-field variational inference algorithms for both models. The generative model is considerably improved by our variational algorithm compared with the previously reported state-of-the-art for that model.

2 Previous Work

Dawid and Skene (1979) laid the groundwork for modern annotation aggregation by proposing the *item-response* model: a probabilistic crowdsourcing model $p(y,a|\gamma)$ over document labels y and annotations a parameterized by confusion matrices γ for each annotator. A growing body of work extends this model to account for such things as correlation among annotators, annotator trustworthiness, item difficulty, and so forth (Bragg et al., 2013; Hovy et al., 2013; Passonneau and Carpenter, 2013; Pasternack and Roth, 2010; Smyth et al., 1995; Welinder et al., 2010; Whitehill et al., 2009; Zhou et al., 2012).

Of the crowdsourcing models that are data-aware, most model the data discriminatively (Carroll et al., 2007; Liu et al., 2012; Raykar et al., 2010; Yan et al., 2014). A smaller line of work models the data generatively (Lam and Stork, 2005; Simpson and Roberts, In Press). We are aware of no papers that compare a generative crowdsourcing model with a similar discriminative model. In the larger context of supervised machine learning, Ng and Jordan (2001) observe that generative models parameters tend to converge with fewer training examples than their discriminatively trained counterparts, but to lower asymptotic performance levels. This paper explores those insights in the context of crowdsourcing models.

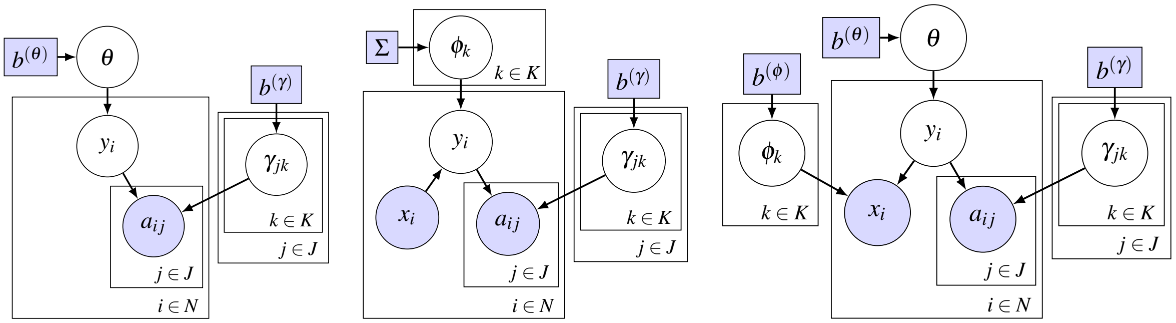
3 Models

At a minimum, a probabilistic crowdsourcing model predicts ground truth labels y from imperfect annotations a (i.e., $\text{argmax}_y\, p(y|a)$). In this section we review the specifics of two previously-proposed data-aware crowdsourcing models. These models are best understood as extensions to a Bayesian formulation of the item-response model that we will refer to as ITEMRESP. ITEMRESP, illustrated in Figure 1a, is defined by the joint distribution

$$\begin{aligned} &p(\theta,\gamma,y,a) \qquad (1)\\ &= p(\theta)\Big[\prod_{j\in J}\prod_{k\in K} p(\gamma_{jk})\Big]\prod_{i\in N} p(y_i|\theta)\prod_{j\in J} p(a_{ij}|\gamma_j,y_i)\end{aligned}$$

where J is the set of annotators, K is the set of class labels, N is the set of data instances in the corpus, θ is a stochastic vector in which θ_k is the probability of label class k, γ_j is a matrix of stochastic vector rows in which $\gamma_{jkk'}$ is the probability that annotator j annotates with k' items whose true label is k, y_i is the class label associated with the ith instance in the corpus, and a_{ijk} is the number of times that instance i was annotated by annotator j with label k. The fact that a_{ij} is a count vector allows for the general case where annotators express their uncertainty over multiple class values. Also, $\theta \sim Dirichlet(b^{(\theta)})$, $\gamma_{jk} \sim Dirichlet(b^{(\gamma)}_{jk})$, $y_i|\theta \sim Categorical(\theta)$, and $a_{ij}|y_i,\gamma_j \sim Multinomial(\gamma_{jy_i},M_i)$ where M_i is the number of times annotator j annotated instance i. We need not define a distribution over M_i because in practice $M_i = |a_{ij}|_1$ is fixed and known during posterior inference. A special case of this model formulates a_{ij} as a categorical distribution assuming that annotators will provide at most one annotation per item. All hyperparameters are designated b and are disambiguated with a superscript (e.g., the hyperparameters for $p(\theta)$ are $b^{(\theta)}$). When ITEMRESP parameters are set with uniform θ values and diagonal confusion matrices γ, majority vote is obtained.

Inference in a crowdsourcing model involves a corpus with an annotated portion $N^A = \{i : |a_i|_1 > 0\}$ and also potentially an unannotated portion $N^U = \{i : |a_i|_1 = 0\}$. ITEMRESP can be written as $p(\gamma,y,a) = p(\gamma,y^A,y^U,a)$ where $y^A = \{y_i : i \in N^A\}$ and $y^U = \{y_i : i \in N^U\}$. However, because ITEMRESP has no model of the data x, it receives no benefit from unannotated data N^U.



(a) ItemResp (b) LogResp (c) MomResp

Figure 1: Directed graphical model depictions of the models discussed in this paper. Round nodes are variables with distributions. Rectangular nodes are hyperparameters (without distributions). Shaded nodes have known values (although some a values may be unobserved).

3.1 Log-linear data model (LogResp)

One way to make ItemResp data-aware is by adding a discriminative log-linear data component (Raykar et al., 2010; Liu et al., 2012). For short, we refer to this model as LogResp, illustrated in Figure 1b. Concretely,

$$p(\gamma,\phi,y,a|x) = \Big[\prod_{j\in J}\prod_{k\in K} p(\gamma_{jk})\Big] \prod_{k\in K} p(\phi_k)\prod_{i\in N} p(y_i|x_i,\phi)\prod_{j\in J} p(a_{ij}|\gamma_j,y_i) \tag{2}$$

where x_{if} is the value of feature f in data instance i (e.g., a word count in a text classification problem), ϕ_{kf} is the probability of feature f occurring in an instance of class k, $\phi_k \sim Normal(0,\Sigma)$, and $y_i|x_i,\phi \sim LogLinear(x_i,\phi)$. That is, $p(y_i|x_i,\phi) = \exp[\phi_{y_i}^T x_i]/\sum_k \exp[\phi_k^T x_i]$.

In the special case that each γ_j is the identity matrix (each annotator is perfectly accurate), LogResp reduces to a multinomial logistic regression model. Because it is a conditional model, LogResp lacks any built-in capacity for semi-supervised learning.

3.2 Multinomial data model (MomResp)

An alternative way to make ItemResp data-aware is by adding a generative multinomial data component (Lam and Stork, 2005; Felt et al., 2014). We refer to the model as MomResp, shown in Figure 1c.

$$p(\theta,\gamma,\phi,y,x,a) = p(\theta)\Big[\prod_{j\in J}\prod_{k\in K} p(\gamma_{jk})\Big] \prod_{k\in K} p(\phi_k)\prod_{i\in N} p(y_i|\theta)p(x_i|y_i,\phi)\prod_{j\in J} p(a_{ij}|\gamma_j,y_i) \tag{3}$$

where ϕ_{kf} is the probability of feature f occurring in an instance of class k, $\phi_k \sim Dirichlet(b_k^{(\phi)})$, $x_i \sim Multinomial(\phi_{y_i},T_i)$, and T_i is a number-of-trials parameter (e.g., for text classification T_i is the number of words in document i). $T_i = |x_i|_1$ is observed during posterior inference $p(\theta,\gamma,\phi,y|x,a)$.

Because MomResp is fully generative over the data features x, it naturally performs semi-supervised learning as data from unannotated instances N^U inform inferred class labels y^A of annotated instances via ϕ. This can be seen by observing that $p(x)$ terms prevent terms involving y^U from summing out of the marginal distribution $p(\theta,\gamma,\phi,y^A,x,a) = \sum_{y^U} p(\theta,\gamma,\phi,y^A,y^U,x,a) = p(\theta,\gamma,\phi,y^A,x^A,a)\sum_{y^U} p(y^U|\theta)p(x^U|y^U)$.

When $N = N^U$ (the unsupervised setting) the posterior distribution $p(\theta,\gamma,\phi,y^U|x,a) = p(\theta,\phi,y^U|x)$ is a mixture of multinomials clustering model. Otherwise, the model resembles a semi-supervised naïve Bayes classifier (Nigam et al., 2006). However, naïve Bayes is supervised by trustworthy labels whereas MomResp is supervised by imperfect annotations mediated by inferred annotator error characteristic γ. In the special case that γ is the identity matrix (each annotator is perfectly accurate), MomResp reduces to a possibly semi-supervised naïve

Bayes classifier where each annotation is a fully trusted label.

3.3 A Generative-Discriminative Pair

MOMRESP and LOGRESP are a generative-discriminative pair, meaning that they belong to the same parametric model family but with parameters fit to optimize joint likelihood and conditional likelihood, respectively. This relationship is seen via the equivalence of the conditional probability of LOGRESP $p_L(y,a|x)$ and the same expression according to MOMRESP $p_M(y,a|x)$. For simplicity in this derivation we omit priors and consider ϕ, θ, and γ to be known values. Then

$$\begin{aligned} p_M(y,a|x) &= \frac{p(y)p(x|y)p(a|y)}{\sum_{y'}\sum_{a'} p(y')p(x|y')p(a'|y')} && (4) \\ &= \frac{p(y)p(x|y)}{\sum_{y'} p(y')p(x|y')} \cdot p(a|y) && (5) \\ &= \frac{\exp[e^{w_y^T x+z}]}{\sum_k \exp[e^{w_k^T x+z}]} \cdot p(a|y) && (6) \\ &= p_L(y,a|x) && (7) \end{aligned}$$

Equation 4 follows from Bayes Rule and conditional independence in the model. In Equation 5 $p(a'|y)$ sums to 1. The first term of Equation 6 is the posterior $p(y|x)$ of a naïve Bayes classifier, known to have the same form as a logistic regression classifier where parameters w and z are constructed from ϕ and θ.[2]

4 Mean-field Variational Inference (MF)

In this section we present novel mean-field (MF) variational algorithms for LOGRESP and MOMRESP. Note that Liu et al. (2012) present (in an appendix) variational inference for LOGRESP based on belief propagation (BP). They do not test their algorithm for LOGRESP; however, their comparison of MF and BP variational inference for the ITEMRESP model indicates that the two flavors of variational inference perform very similarly. Our MF algorithm for LOGRESP has not been designed with the idea of outperforming its BP analogue, but rather with the goal of ensuring that the generative and discriminative model use the same inference algorithm. We expect that we would achieve the same results if our comparison used variational BP algorithms for both MOMRESP and LOGRESP, although such an additional comparison is beyond the scope of this work.

Broadly speaking, variational approaches to posterior inference transform inference into an optimization problem by searching within some family of tractable approximate distributions Q for the distribution $q \in Q$ that minimizes distributional divergence from an intractable target posterior p^*. In particular, under the mean-field assumption we confine our search to distributions Q that are fully factorized.

4.1 LOGRESP Inference

We approximate LOGRESP's posterior $p^*(\gamma,\phi,y|x,a)$ using the fully factorized approximation $q(\gamma,\phi,y) = \left[\prod_j \prod_k q(\gamma_{jk})\right] \prod_k q(\phi_k) \prod_i q(y_i)$. Approximate marginal posteriors q are disambiguated by their arguments.

Algorithm. Initialize each $q(y_i)$ to the empirical distribution observed in the annotations a_i. The Kullback-Leibler divergence $KL(q||p^*)$ is minimized by iteratively updating each variational distribution in the model as follows:

$$q(\gamma_{jk}) \propto \prod_{k' \in K} \gamma_{jkk'}^{b_{jkk'}^{(\gamma)} + \sum_{i\in N} a_{ijk'} q(y_i=k) - 1} = Dirichlet(\alpha_{jk}^{(\gamma)})$$

$$q(\phi_k) \propto \exp\left[\phi_k^T \Sigma^{-1} \phi_k + \sum_{i \in N} q(y_i = k) \phi_k^T x_i\right]$$

$$\begin{aligned} q(y_i) &\propto \prod_{k\in K} \exp\Big[\sum_{j\in J}\sum_{k'\in K} a_{ijk'} E_{q(\gamma_{jk})}[\log \gamma_{jkk'}] + \\ &\qquad \sum_{f \in F} x_{if} E_{q(\phi_k)}[\phi_{kf}]\Big]^{\mathbb{1}(y_i=k)} \\ &\propto \prod_{k\in K} \alpha_{ik}^{(y)\mathbb{1}(y_i=k)} = Categorical(\alpha_i^{(y)}) \end{aligned}$$

Approximate distributions are updated by calculating variational parameters $\alpha^{(\cdot)}$, disambiguated by a superscript. Because $q(\gamma_{jk})$ is a Dirichlet distribution the term $E_{q(\gamma_{jk})}[\log \gamma_{jkk'}]$ appearing in $q(y_i)$ is computed analytically as $\psi(\alpha_{jkk'}^{(\gamma)}) - \psi(\sum_{k'} \alpha_{jkk'}^{(\gamma)})$ where ψ is the digamma function.

The distribution $q(\phi_k)$ is a logistic normal distribution. This means that the expectations $E_{q(\phi_k)}[\phi_{kf}]$ that appear in $q(y_i)$ cannot be computed analytically. Following Liu et al. (2012), we approximate the distribution $q(\phi_k)$ with the point estimate

[2] http://cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf gives a proof of this property in the continuous case and hints about the discrete case proof.

$\hat{\phi}_k = \mathrm{argmax}_{\phi_k}\, q(\phi_k)$ which can be calculated using existing numerical optimization methods for log-linear models. Such maximization can be understood as embedding the variational algorithm inside of an outer EM loop such as might be used to tune hyperparameters in an empirical Bayesian approach (where ϕ are treated as hyperparameters).

4.2 MOMRESP Inference

MOMRESP's posterior $p^*(y,\theta,\gamma,\phi|x,a)$ is approximated with the fully factorized distribution $q(y,\theta,\gamma,\phi) = q(\theta)\left[\prod_j \prod_k q(\gamma_{jk})\right] \prod_k q(\phi_k) \prod_i q(y_i)$.

Algorithm. Initialize each $q(y_i)$ to the empirical distribution observed in the annotations a_i. The Kullback-Leibler divergence $KL(q||p^*)$ is minimized by iteratively updating each variational distribution in the model as follows:

$$q(\theta) \propto \prod_{k \in K} \theta_k^{b_k^{(\theta)} + \sum_{i \in N} q(y_i = k) - 1} = Dirichlet(\alpha^{(\theta)})$$

$$q(\gamma_{jk}) \propto \prod_{k' \in K} \gamma_{jkk'}^{b_{jkk'}^{(\gamma)} + \sum_{i \in N} a_{ijk'} q(y_i = k) - 1} = Dirichlet(\alpha_{jk}^{(\gamma)})$$

$$q(\phi_k) \propto \prod_{f \in F} \phi_{kf}^{b_{kf}^{(\phi)} + \sum_{i \in N} x_{if} q(y_i = k) - 1} = Dirichlet(\alpha_k^{(\phi)})$$

$$\begin{aligned} q(y_i) \propto \prod_{k \in K} \exp\Big[\sum_{j \in J} \sum_{k' \in K} a_{ijk'} E_{q(\gamma_{jk})}[\log \gamma_{jkk'}] + \\ E_{q(\theta_k)}[\log \theta_k] + \sum_{f \in F} x_{if} E_{q(\phi_k)}[\log \phi_{kf}] \Big]^{\mathbb{1}(y_i = k)} \\ \propto \prod_{k \in K} \alpha_{ik}^{(y) \mathbb{1}(y_i = k)} = Categorical(\alpha_i^{(y)}) \end{aligned}$$

Approximate distributions are updated by calculating the values of variational parameters $\alpha^{(\cdot)}$, disambiguated by a superscript. The expectations of log terms in the $q(y_i)$ update are all with respect to Dirichlet distributions and so can be computed analytically as explained previously.

4.3 Model priors and implementation details

Computing a lower bound on the log likelihood shows that in practice the variational algorithms presented above converge after only a dozen or so updates. We compute $\mathrm{argmax}_{\phi_k}\, q(\phi_k)$ for LOGRESP using the L-BFGS algorithm as implemented in MALLET (McCallum, 2002). We choose uninformed priors $b_k^{(\theta)} = 1$ for MOMRESP and identity matrix $\Sigma = \mathbb{1}$ for LOGRESP. We set $b_{kf}^{(\phi)} = 0.1$ for MOMRESP to encourage sparsity in per-class word distributions. Liu et al. (2012) argue that a uniform prior over the entries of each confusion matrix γ_j can lead to degenerate performance. Accordingly, we set the diagonal entries of each $b_j^{(\gamma)}$ to a higher value $b_{jkk}^{(\gamma)} = \frac{1+\delta}{K+\delta}$ and off-diagonal entries to a lower value $b_{jkk'}^{(\gamma)} = \frac{1}{K+\delta}$ with $\delta = 2$.

Both MOMRESP and LOGRESP are given full access to all instances in the dataset, annotated and unannotated. However, as explained in Section 3.1, LOGRESP is conditioned on the data and thus is structurally unable to make use of unannotated data. We experimented briefly with self-training for LOGRESP but it had little effect. With additional effort one could likely settle on a heuristic scheme that allowed LOGRESP to benefit from unannotated data. However, since such an extension is external to the model itself, it is beyond the scope of this work.

5 Experiments with Simulated Annotators

Models which learn from error-prone annotations can be challenging to evaluate in a systematic way. Simulated annotations allow us to systematically control annotator behavior and measure the performance of our models in each configuration.

5.1 Simulating Annotators

We simulate an annotator by corrupting ground truth labels according to that annotator's accuracy parameters. Simulated annotators are drawn from the annotator quality pools listed in Table 1. Each row is a named pool and contains five annotators A1–A5, each with a corresponding accuracy parameter (the number five is chosen arbitrarily). In the pools HIGH, MED, and LOW, annotator errors are distributed uniformly across the incorrect classes. Because there are no patterns among errors, these settings approximate situations in which annotators are ultimately in agreement about the task they are doing, although some are better at it than others. The HIGH pool represents a corpus annotation project with high quality annotators. In the MED and LOW pools annotators are progressively less reliable.

The CONFLICT annotator pool in Table 1 is special in that annotator errors are made systematically rather than uniformly. Systematic errors are

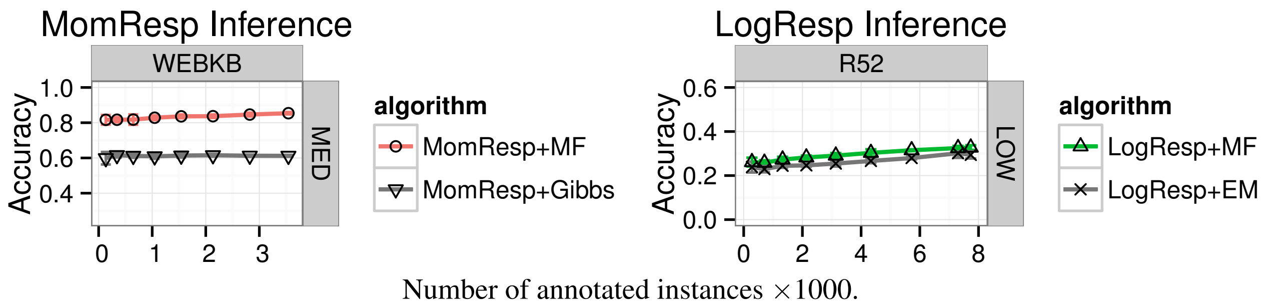


Figure 2: Mean field (MF) variational inference outperforms previous inference methods for both models. *Left*: MomResp with MF (MomResp+MF) versus with Gibbs sampling (MomResp+Gibbs) on the WebKB dataset using annotators from the MED pool. *Right*: LogResp with MF (LogResp+MF) versus with EM (LogResp+EM) on the Reuters52 dataset using annotators from the LOW pool.

produced at simulation time by constructing a per-annotator confusion matrix (similar to γ_j) whose diagonal is set to the desired accuracy setting, and whose off-diagonal row entries are sampled from a symmetric Dirichlet distribution with parameter 0.1 to encourage sparsity and then scaled so that each row properly sums to 1. These draws from a sparse Dirichlet yield consistent error patterns. The CONFLICT pool approximates an annotation project where annotators understand the annotation guidelines differently from one another. For the sake of example, annotator A5 in the CONFLICT setting will annotate documents with the true class B as B exactly 10% of the time but might annotate B as C 85% of the time. On the other hand, annotator A4 might annotate B as D most of the time. We choose low agreement rates for CONFLICT to highlight a case that violates majority vote's assumption that annotators are basically in agreement.

	A1	A2	A3	A4	A5
HIGH	90	85	80	75	70
MED	70	65	60	55	50
LOW	50	40	30	20	10
CONFLICT	50†	40†	30†	20†	10†

Table 1: For each simulated annotator quality pool (HIGH, MED, LOW, CONFLICT), annotators A1-A5 are assigned an accuracy. † indicates that errors are systematically in conflict as described in the text.

5.2 Datasets and Features

We simulate the annotator pools from Table 1 on each of six text classification datasets. The datasets 20 Newsgroups, WebKB, Cade12, Reuters8, and Reuters52 are described by Cardoso-Cachopo (2007). The LDC-labeled Enron emails dataset is described by Berry et al. (2001). Each dataset is preprocessed via Porter stemming and by removal of the stopwords from MALLET's stopword list. Features occurring fewer than 5 times in the corpus are discarded. Features are fractionally scaled so that $|x_i|_1$ is equal to the average document length since document scaling has been shown to be beneficial for multinomial document models (Nigam et al., 2006).

Each dataset is annotated according to the following process: an instance is selected at random (without replacement) and annotated by three annotators selected at random (without replacement). Because annotation simulation is a stochastic process, each simulation is repeated five times.

5.3 Validating Mean-field Variational Inference

Figure 2 compares mean-field variational inference (MF) with alternative inference algorithms from previous work. For variety, the left and right plots are calculated over arbitrarily chosen datasets and annotator pools, but these trends are representative of other settings. MomResp using MF is compared with MomResp using Gibbs sampling estimating $p(y|x,a)$ from several hundred samples (an improvement to the method used by Felt et al. (2014)).

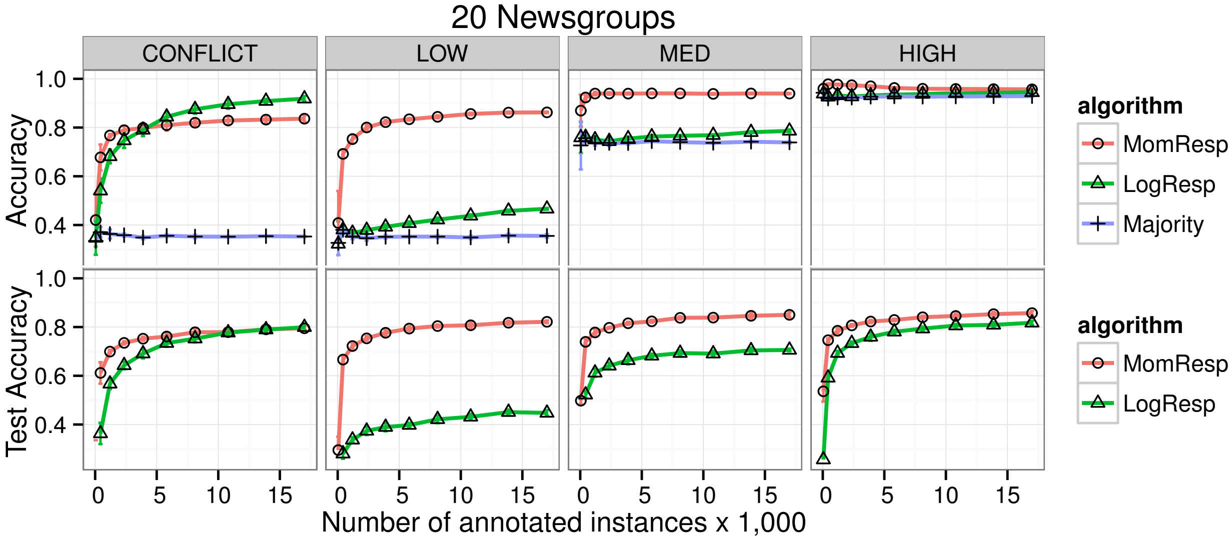


Figure 3: *Top row*: Inferred label accuracy on three-deep annotations. A majority vote baseline is shown for reference. *Bottom row*: Generalization accuracy on a test set. Majority vote is not shown since it does not generate test set predictions. Each column uses the indicated simulated annotator pool.

MOMRESP benefits significantly from MF. We suspect that this disparity could be reduced via hyperparameter optimization as indicated by Asuncion et al. (2009). However, that investigation is beyond the scope of the current work. LOGRESP using MF is compared with LOGRESP using expectation maximization (EM) as in (Raykar et al., 2010). LOGRESP with MF displays minor improvements over LOGRESP with EM. This is consistent with the modest gains that Liu et al. (2012) reported when comparing variational and EM inference for the ITEMRESP model.

5.4 Discriminative (LOGRESP) versus Generative (MOMRESP)

We run MOMRESP and LOGRESP with MF inference on the cross product of datasets and annotator pools. Inferred label accuracy on items that have been annotated is the primary task of crowdsourcing; we track this measure accordingly. However, the ability of these models to generalize on unannotated data is also of interest and allows better comparison with traditional non-crowdsourcing models. Figure 3 plots learning curves for each annotator pool on the 20 Newsgroups dataset; results on other datasets are summarized in Table 2. The first row of Figure 3 plots the accuracy of labels inferred from annotations. The second row of Figure 3 plots generalization accuracy using the inferred model parameters ϕ (and θ in the case of MOMRESP) on held-out test sets with no annotations. The generalization accuracy curves of MOMRESP and LOGRESP may be compared with those of naïve Bayes and logistic regression, respectively. Recall that in the special case where annotations are both flawless and trusted (via diagonal confusion matrices γ) then MOMRESP and LOGRESP simplify to semi-supervised naïve Bayes and logistic regression classifiers, respectively.

Notice that MOMRESP climbs more steeply than LOGRESP in all cases. This observation is in keeping with previous work in supervised learning. Ng and Jordan (2001) argue that generative and discriminative models have complementary strengths: generative models tend to have steeper learning curves and converge in terms of parameter values after only $\log n$ training examples, whereas discriminative models tend to achieve higher asymptotic levels but converge more slowly after n training examples. The second row of Figure 3 shows that even after training on three-deep annotations over the entire 20 newsgroups dataset, LOGRESP's data model does not approach its asymptotic level of performance. The early steep slope of the generative model is more desirable in this setting than the eventually superior performance of the discriminative model given large numbers of annotations. Figure 4 additionally plots MomRespA, a variant of MomResp deprived of all unannotated documents, showing that the early generative advantage is not attributable entirely to semi-supervision.

The generative model is more robust to annotation noise than the discriminative model, seen by comparing the LOW, MED, and HIGH columns in Figure 3. This robustness is significant because crowdsourcing tends to yield noisy annotations, making the LOW and MED annotator pools of greatest practical interest. This assertion is borne out by an experiment with CrowdFlower, reported in Section 6.

To validate that LogResp does, indeed, asymptotically surpass MomResp we ran inference on datasets with increasing annotation depths. Crossover does not occur until 20 Newsgroups is annotated nearly 12-deep for LOW, 5-deep for MED, and 3.5-deep (on average) for HIGH. Additionally, for each combination of dataset and annotator pool except those involving CONFLICT, by the time LogResp surpasses MomResp, the majority vote baseline is extremely competitive with LogResp. The CONFLICT setting is the exception to this rule: CONFLICT annotators are particularly challenging for majority vote since they violate the implicit assumption that annotators are basically aligned with the truth. The CONFLICT setting is of practical interest only when annotators have dramatic deep-seated differences of opinion about what various labels should mean. For most crowdsourcing projects this issue may be avoided with sufficient up-front orientation of the annotators. For reference, in Figure 4 we show that a less extreme variant of CONFLICT behaves more similarly to LOW.

Table 2 reports the percent of the dataset that must be annotated three-deep before LogResp's inferred label accuracy surpasses that of MomResp. Crossover tends to happen later when annotation quality is low and earlier when annotator quality is high. Cases reported as *NA* were too close to call; that is, the dominating algorithm changed depending on the random run.

Unsurprisingly, MomResp is not well suited to all classification datasets. The 0% entries in Table 2 mean that LogResp dominates the learning curve for that annotator pool and dataset. These cases are likely the result of the MomResp model making the same strict inter-feature independence assumptions as naïve Bayes, rendering it tractable and effective for many classification tasks but ill-suited for datasets where features are highly correlated or for tasks in which class identity is not informed by document vocabulary. The CADE12 dataset, in particular, is known to be challenging. A supervised naïve Bayes classifier achieves only 57% accuracy on this dataset (Cardoso-Cachopo, 2007). We would expect MomResp to perform similarly poorly on sentiment classification data. Although we assert that generative models are inherently better suited to crowdsourcing than discriminative models, a sufficiently strong mismatch between model assumptions and data can negate this advantage.

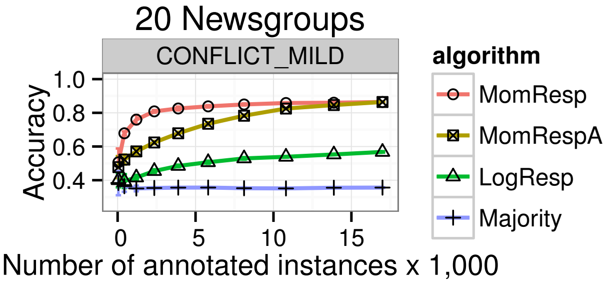


Figure 4: Inferred label accuracy for a variant of the CONFLICT annotator pool in which the off-diagonals of each annotator confusion matrix are drawn from a Dirichlet parameterized by 1 rather than 0.1. Also adds the algorithm MomRespA to show the effect of removing MomResp's access to unannotated documents.

6 Experiments with Human Annotators

In the previous section we used simulations to control annotator error. In this section we relax that control. To assess the effect of real-world annotation error on MomResp and LogResp, we selected 1000 instances at random from 20 Newsgroups and paid annotators on CrowdFlower to annotate them with the 20 Newsgroups categories, presented as human-readable names (e.g., "Atheism" for alt.atheism). Annotators were allowed to express uncertainty by

	CONFLICT	LOW	MED	HIGH
20 News	21%	✓	✓	✓
WebKB	*NA*	✓	✓	0%
Reuters8	*NA*	✓	✓	✓
Reuters52	✓	✓	✓	✓
CADE12	0%	✓	0%	0%
Enron	✓	✓	✓	18%

Table 2: The percentage of the dataset that must be annotated (three-deep) before the generative model MOMRESP is surpassed by LOGRESP. ✓indicates that MOMRESP dominates the entire learning curve; 0% indicates that LOGRESP dominates. *NA* indicates high variance cases that were too close to call.

selecting up to three unique categories per document. During the course of a single day we gathered 7,265 annotations, with each document having a minimum of 3 and a mean of 7.3 annotations.[3] Figure 5 shows learning curves for the CrowdFlower annotations. The trends observed previously are unchanged. MOMRESP enjoys a significant advantage when relatively few annotations are available. Presumably LOGRESP would still dominate if we were able to explore later portions of the curve or curves with greater annotation depth.

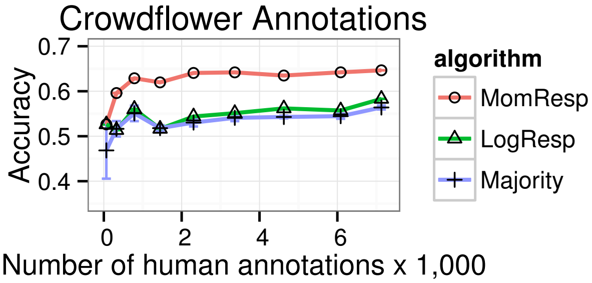


Figure 5: Inferred label accuracy on annotations gathered from CrowdFlower over a subset of 1000 instances of the 20 Newsgroups dataset. At the last plotted point there are $7{,}265/1{,}000 \approx 7.3$ annotations per instance.

7 Conclusions and Future Work

We have argued that generative models are better suited than discriminative models to the task of annotation aggregation since they tend to be more robust to annotation noise and to approach their asymptotic performance levels with fewer annotations. Also, in settings where a discriminative model would usually shine, there are often enough annotations that a simple baseline of majority vote is sufficient.

In support of this argument, we developed comparable mean-field variational inference for a generative-discriminative pair of crowdsourcing models and compared them on both crowdsourced and synthetic annotations on six text classification datasets. In practice we found that on classification tasks for which generative models of the data work reasonably well, the generative model greatly outperforms its discriminative log-linear counterpart.

The generative multinomial model we employed makes inter-feature independence assumptions ill suited to some classification tasks. Document topic models (Blei, 2012) could be used as the basis of a more sophisticated generative crowdsourcing model. One might also transform the data to make it more amenable to a simple model using documents assembled from distributed word representations (Mikolov et al., 2013). Finally, although we expect these results to generalize, we have only experimented with text classification. Similar experiments could be performed on other commonly crowdsourced tasks such as sequence labeling.

Acknowledgments

We thank Alex Smola and the anonymous reviewers for their insightful comments. This work was supported by the collaborative NSF Grant IIS-1409739 (BYU) and IIS-1409287 (UMD).

References

[Asuncion et al.2009] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press.

[Berry et al.2001] M. W. Berry, M. Browne, and

[3] This dataset and the scripts that produced it are available via git at git://nlp.cs.byu.edu/plf1/crowdflower-newsgroups.git

B. Signer. 2001. Topic annotated Enron email data set. *Linguistic Data Consortium, Philadelphia*.

[Blei2012] D. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

[Bragg et al.2013] J. Bragg, Mausam, and D. Weld. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[Cardoso-Cachopo2007] A. Cardoso-Cachopo. 2007. *Improving Methods for Single-label Text Categorization*. Ph.D. thesis, Universidade Tecnica de Lisboa.

[Carroll et al.2007] J. Carroll, R. Haertel, P. McClanahan, E. Ringger, and K. Seppi. 2007. Modeling the annotation process for ancient corpus creation. In *Proceedings of ECAL 2007*, pages 25–42. Charles University.

[Dawid and Skene1979] A.P. Dawid and A.M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28.

[Felt et al.2014] P. Felt, R. Haertel, E. Ringger, and K. Seppi. 2014. MomResp: A Bayesian model for multi-annotator document labeling. In *Proceedings of LREC*.

[Hovy et al.2013] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of HLT-NAACL 2013*, pages 1120–1130.

[Jurgens2013] D. Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of NAACL-HLT 2013*, pages 556–562.

[Lam and Stork2005] C. P. Lam and D. G. Stork. 2005. Toward optimal labeling strategy under multiple unreliable labelers. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*.

[Liu et al.2012] Q. Liu, J. Peng, and A. Ihler. 2012. Variational inference for crowdsourcing. In *NIPS*, pages 692–700.

[McCallum2002] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

[Mikolov et al.2013] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

[Ng and Jordan2001] A. Ng and M. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *NIPS*, 14:841–848.

[Nigam et al.2006] K. Nigam, A. McCallum, and T. Mitchell. 2006. Semi-supervised text classification using EM. *Semi-Supervised Learning*, pages 33–56.

[Passonneau and Carpenter2013] R. Passonneau and B. Carpenter. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195. Citeseer.

[Pasternack and Roth2010] J. Pasternack and D. Roth. 2010. Knowing what to believe (when you already know something). In *COLING*, Beijing, China.

[Raykar et al.2010] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy. 2010. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322.

[Simpson and RobertsIn Press] E. Simpson and S. Roberts. In Press. Bayesian methods for intelligent task assignment in crowdsourcing systems. In *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*. Springer.

[Smyth et al.1995] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. 1995. Inferring ground truth from subjective labelling of Venus images. *NIPS*, pages 1085–1092.

[Snow et al.2008] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*. ACL.

[Surowiecki2005] J. Surowiecki. 2005. *The Wisdom of Crowds*. Random House LLC.

[Welinder et al.2010] P. Welinder, S. Branson, P. Perona, and S. Belongie. 2010. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432.

[Whitehill et al.2009] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *NIPS*, 22:2035–2043.

[Yan et al.2014] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy. 2014. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327.

[Zhou et al.2012] D. Zhou, J. Platt, S. Basu, and Y. Mao. 2012. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, volume 25, pages 2204–2212.