A Bayesian Framework for Modeling Human Evaluations

Himabindu Lakkaraju* Jure Leskovec* Jon Kleinberg† Sendhil Mullainathan‡

Abstract

Several situations that we come across in our daily lives involve some form of *evaluation*: a process where an *evaluator* chooses a correct label for a given *item*. Examples of such situations include a crowd-worker labeling an image or a student answering a multiple-choice question. Gaining insights into human evaluations is important for determining the quality of individual evaluators as well as identifying true labels of items. Here, we generalize the question of estimating the quality of individual evaluators, extending it to obtain diagnostic insights into how various evaluators label different kinds of items. We propose a series of increasingly powerful hierarchical Bayesian models which infer latent groups of evaluators and items with the goal of obtaining insights into the underlying evaluation process. We apply our framework to a wide range of real-world domains, and demonstrate that our approach can accurately predict evaluator decisions, diagnose types of mistakes evaluators tend to make, and infer true labels of items.

1 Introduction

Several seemingly unrelated tasks such as a crowd-worker on Amazon Mechanical Turk labeling an image, a librarian classifying a newly arrived title, a Yelp user rating a restaurant, or a student providing an answer to a multiple-choice test share an underlying theme. All of these and many more such situations are examples of human evaluation processes, in which an *evaluator* is shown an *item* and attempts to choose a *correct label* for it.

The result of each such evaluation depends on the attributes of both the evaluator and the item. For example, consider a crowd-worker (the evaluator) labeling images of birds (the items). The quality of the labels produced will likely depend on the characteristics of the crowd-worker, including her general level of expertise about birds and/or her experience with different geographical regions; the quality will also depend on the characteristics of the birds being labeled.

*Stanford University, {himalv,jure}@cs.stanford.edu
†Cornell University, kleinber@cs.cornell.edu
‡Harvard University, mullain@fas.harvard.edu

A long line of research has studied how to take multiple labels from non-expert evaluators and synthesize them into a single high-quality label [1]-[7], and how to estimate the performance of individual evaluators on various tasks [8]-[12]. However, relatively little attention has been focused on obtaining deeper insights into evaluations such as understanding the characteristics of mistakes being made and identifying the shared attributes of evaluators and items that are relevant to the quality of the resulting label. Discovering these patterns may in turn generate diagnostic insights such as which kinds of items are particularly hard to label or what types of mistakes certain kinds of evaluators are making.

In order to understand human evaluations, Dawid and Skene [10] proposed a model for estimating *confusion matrices* of individual evaluators. A *confusion matrix* models the labeling decisions of an evaluator. In a confusion matrix $\Theta^{(j)}$, entry (p, q) is the probability of an item with *true* label p being assigned label q by an evaluator j. Error-free evaluation corresponds to a diagonal confusion matrix, while off-diagonal entries record different types of errors. However, the problem is that often it is too expensive to obtain enough evaluations and enough ground-truth labels to estimate a separate confusion matrix for each evaluator. Further, it might not be possible to explain all the decisions of an evaluator with just one such confusion matrix. This is due to the fact that decisions also depend upon the characteristics of the items that an evaluator is judging.

Present work: Obtaining diagnostic insights into human evaluations. In this work, we provide a framework for obtaining insights into human evaluations by casting it as a problem of inferring confusion matrices which explain the decisions made by evaluators.

In order to address the aforementioned drawbacks of existing solutions, we propose a novel hierarchical Bayesian framework. The crux of this framework involves inferring two sets of clusters - groups of evaluators and items respectively - which can guide the process of estimating the confusion matrices. The intuition behind the clustering process is to group together all the evaluators who share similar attributes and evaluation styles. Similarly, all the items grouped into the same cluster would share similar attributes and are likely to

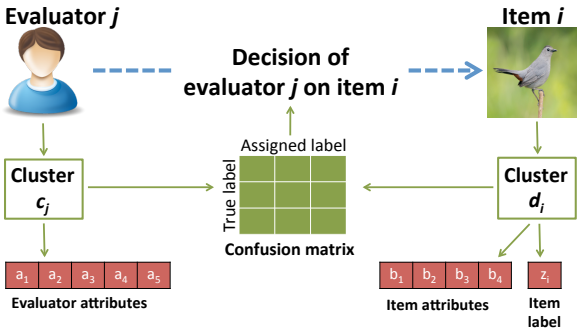


Figure 1: Overview of our approach.

be subjected to similar kinds of evaluations. Figure 1 illustrates our approach. Each evaluator j is assigned to a latent cluster c_j and each item i to a latent cluster d_i. Each such cluster pair (c_j, d_i) is associated with a latent confusion matrix which captures how evaluators of cluster c_j label items of cluster d_i.

Our framework allows for inferring true labels of items, predicting labels given by evaluators, and also recovering evaluator confusion matrices. Our approach identifies sub-populations of evaluators and the common properties of the items on which they make mistakes. For example, consider students answering multiple-choice questions; using our methodology we might find that students taking a machine learning class naturally cluster into those with a background in statistics and those with a background in optimization, and that the statistics students perform better on questions pertaining to graphical models, while students with a background in optimization perform better on Support Vector Machines questions. Note that such insights arise only when we simultaneously consider evaluators (students), their attributes (statistics vs. optimization), items (questions) and their attributes (graphical models vs. SVMs) in the modeling process.

We demonstrate the generality of our approach by applying it across a number of different domains: image and text labeling tasks, students answering multiple-choice questions, and peer grading of course assignments. For each of these domains, our models outperform state-of-the-art approaches on a variety of tasks such as predicting evaluator decisions, inferring true labels of items, and estimating evaluator confusion matrices. We also obtain interesting domain-specific insights and identify several interesting patterns of evaluations.

2 Proposed Approach

In this section, we propose a series of hierarchical Bayesian models for obtaining diagnostic insights into how evaluators label different types of items. Table 1 summarizes the notation.

Background. Our probabilistic framework is inspired by a simple and elegant approach for capturing the quality as well as the error properties of individual evaluators due to Dawid and Skene [10]. The underlying assumption of this model is that each item has a true label, and an evaluator's prediction of the item's label is a function of the true label masked by the evaluator's own confusion matrix. Assuming that there are K labels an item can have, the model associates a confusion matrix $\Theta^{(j)}$, of size $K \times K$, with each evaluator j. Each element (p, q) of an evaluator's confusion matrix then represents the probability that the evaluator labels an item with label q when the true label of the item is p. Table 2 describes the generative process for the Dawid-Skene model.

Symbol	Description
J	Number of evaluators
I	Number of items
K	Number of the classes of items
i	Index for items
j	Index for evaluators
$r_{i,j}$	Label assigned by evaluator j to item i
N	Number of attributes of evaluators
M	Number of attributes of items
$a_n^{(j)}$	n^{th} attribute of evaluator j
$b_m^{(i)}$	m^{th} attribute of item i
L_1	Number of clusters of evaluators
L_2	Number of clusters of items
$\Theta^{(j)}$	Confusion matrix of evaluator j
c_j	Cluster label assigned to evaluator j
d_i	Cluster label assigned to item i
$\Theta'^{(c)}$	Confusion matrix of cluster c
z_i	True label of item i
ρ, α, α', β, β'	Model parameters
ϵ_ρ, ϵ_α, ϵ_β, $\epsilon_{\alpha'}$, $\epsilon_{\beta'}$, $\wedge$	Hyperparameters and priors

Table 1: Definition of symbols and concepts.

For each item i,
Sample the item label $z_i \sim$ Multinomial(ρ)

For each decision involving evaluator j and item i,
Sample the decision $r_{i,j} \sim$ Multinomial$(\Theta^{(j)}_{z_i,:})$

Table 2: Generative process for Dawid-Skene model.

Our Framework. A major difference between our framework and previous approaches [10, 12] is that our framework can obtain deeper insights into the evaluation process in addition to inferring the error properties of evaluators, and estimating the true labels of items. In order to obtain such insights, we model evaluators and items as entities with *attributes*. For instance, when we model the process of a student answering a question on an exam, we consider a student as an entity with various attributes such as his expertise, personal characteris-

tics, and cognitive abilities. By adding attributes to the modeling framework, we can connect these attributes with the evaluations themselves, thus facilitating the process of finding interesting patterns in the evaluation process.

Next, we propose three models: *Evaluator Confusion* provides a means to discover patterns among evaluators by grouping them based on their attributes and behaviors; *Item Confusion* analogously groups items; Finally, *Joint Confusion* is designed to discover interactions between the attributes of evaluators and items. We conclude by presenting nonparametric versions of each of these models which can readily accomodate changes in the populations of evaluators and items. Our framework is designed to find a balance between the number of free parameters in the model and its flexibility.

Evaluator Confusion (EC) model. The Evaluator Confusion model groups together evaluators who share similar attributes and also exhibit similar patterns of decision making. The model then assigns a single confusion matrix to each such group. This lets us study shared attributes of evaluators who have similar decision making styles, and the characteristics of their evaluations. For instance, when modeling student performance on a machine learning exam, this model can provide insights such as *students with little expertise in linear algebra have a higher probability of making a mistake.*

More precisely, in the Evaluator Confusion model each item i is associated with a hidden label z_i and a set of observed characteristics/attributes $b^{(i)}$. Each evaluator j is assigned to a latent cluster $c_j \in \{1 \cdots L_1\}$ such that c_j governs the decisions made by evaluator j as well as the evaluator's attributes $a^{(j)}$. Therefore, $\Theta^{(j)} = \Theta'^{(c_j)}$ for every evaluator j assigned to cluster c_j. Now, each decision made by j is an interaction between an item i with label z_i and the cluster c_j. The decisions made by evaluator j and the evaluator's attributes are observed. Further, we assume that all of the latent confusion matrices are sampled from a common Dirichlet prior $\wedge$. We also assume Dirichlet priors on the other multinomial distributions. The generative process is outlined in Table 3.

Item Confusion (IC) model. The Item Confusion model groups items rather than evaluators, and associates a shared confusion matrix with each group of items. This lets us study shared attributes of items on which evaluators make similar mistakes, as well as the shared confusion matrices these items generate. For example, this model can discover patterns such as *students have a higher probability of making a mistake on questions pertaining to advanced topics in the course.*

For each evaluator j,
  Sample the cluster label $c_j \sim$ Multinomial(α)
  For each of the evaluator's attributes $a_n \in \{1 \cdots N\}$
    Sample the attribute value $a_n^{(j)} \sim$ Multinomial(β_{c_j,a_n})

For each item i,
  Sample the item label $z_i \sim$ Multinomial(ρ)
  For each of the attributes $b_m \in \{1 \cdots M\}$
    Sample the attribute value $b_m^{(i)} \sim$ Multinomial(β'_{b_m})

For each decision involving evaluator j and item i,
  Sample the decision $r_{i,j} \sim$ Multinomial($\Theta'^{(c_j)}_{z_i,:}$)

Table 3: Evaluator Confusion (EC) model.

For each item i,
  Sample the cluster label $d_i \sim$ Multinomial(α)
  Sample the item label $z_i \sim$ Multinomial(ρ_{d_i})
  For each of the item's attributes $b_m \in \{1 \cdots M\}$
    Sample the attribute value $b_m^{(i)} \sim$ Multinomial(β'_{d_i,b_m})

For each evaluator j,
  For each of the attributes $a_n \in \{1 \cdots N\}$
    Sample the attribute value $a_n^{(j)} \sim$ Multinomial(β_{a_n})

For each decision involving evaluator j and item i,
  Sample the decision $r_{i,j} \sim$ Multinomial($\Theta'^{(d_i)}_{z_i,:}$)

Table 4: Item Confusion (IC) model.

In this model, each item i is assigned to a latent cluster $d_i \in \{1 \cdots L_2\}$. The cluster determines the hidden label of the item, the observed attribute values associated with the item i, and the decisions associated with the item. Each such cluster of items is associated with a single latent confusion matrix which guides the decisions pertaining to these items. There is no evaluator-specific aspect in this model and all the evaluators share the same set of confusion matrices. When a evaluator j labels an item i, the label is generated according to the confusion matrix that only depends on i, $\Theta^{(j)} = \Theta'^{(d_i)}$. We also assume that all the confusion matrices associated with various clusters of items are sampled from a common Dirichlet prior $\wedge$. Similarly, the other multinomial distributions that we come across in the generative process are sampled from the corresponding Dirichlet priors. The complete generative process is summarized in Table 4.

Joint Confusion (JC) model. Finally, we describe the Joint Confusion model, which is designed for identifying patterns involving the interplay between evaluators and items. We introduce two latent variables c_j and d_i which correspond to the cluster assignments of

For each evaluator j,
  Sample the cluster label $c_j \sim \text{Multinomial}(\alpha)$
  For each of the attributes $a_n \in \{1 \cdots N\}$
    Sample the attribute value $a_n^{(j)} \sim \text{Multinomial}(\beta_{c_j,a_n})$

For each item i,
  Sample the cluster label $d_i \sim \text{Multinomial}(\alpha)$
  Sample the item label $z_i \sim \text{Multinomial}(\rho_{d_i})$
  For each of the attributes $b_m \in \{1 \cdots M\}$
    Sample the attribute value $b_m^{(i)} \sim \text{Multinomial}(\beta'_{d_i,b_m})$

For each decision involving evaluator j and item i,
  Sample the decision $r_{i,j} \sim \text{Multinomial}(\Theta'^{(c_j,d_i)}_{z_i,:})$

Table 5: Joint Confusion (JC) model.

evaluator j and item i, respectively. The evaluations involving j and i are modeled as interactions between clusters c_j and d_i. This unified model exploits the interactions between evaluators and items in guiding the clustering process. In the process of inferring the cluster assignments, those evaluators who share similar attributes and patterns of mistakes get grouped into a single cluster; the items are grouped similarly. For instance, the model might group all the students with no background in linear algebra together. Similarly, all the questions pertaining to the topic non-negative matrix factorization might be grouped together. This allows us to come up with the insights pertaining to the interactions between these two groups: *students with no background in linear algebra have a higher probability of making a mistake when answering questions pertaining to the topic non-negative matrix factorization.*

In this model, there are $L_1 \times L_2$ confusion matrices, where L_1 corresponds to the number of clusters of evaluators and L_2 corresponds to the number of clusters of items. For every pair consisting of an evaluator cluster and item cluster, there is an associated latent confusion matrix which is estimated during the inference process. All the decisions involving evaluators assigned to a cluster c and items assigned to a cluster d are governed by the confusion matrix $\Theta'^{(c_j,d_i)}$, and all the confusion matrices share a common Dirichlet prior $\wedge$. The generative process is outlined in Table 5.

Nonparametric Joint Confusion (NJC) model. We now describe how to extend Joint Confusion model to handle some additional challenges. First, in most real-world settings, it is not a prior determined how many clusters we should work with. Ideally, the model should provide a means of generating the number of clusters based on the data and some underlying parameters. Second, we would like the model to incorporate new data as it arrives; in their current form, the addition of data to the models is not straightforward, and would requiring re-running the entire algorithm. To deal with these issues, we propose a nonparametric extension.

For our nonparametric extensions we use *Dirichlet processes*, which are a popular prior in clustering applications where the number of clusters cannot be specified a priori but instead grows with the data size [13]. We change the distribution from which we sample the cluster labels in each of the models (Tables 3, 4, 5). Intuitively, we sample cluster labels as follows: the probability that an evaluator (or an item) is assigned to an existing cluster c is proportional to the number of data points currently in that cluster; and the probability that an evaluator (or an item) is assigned to a brand new cluster is proportional to a parameter γ. This kind of distribution results in the creation of new clusters at a greater rate during the beginning of the estimation process, when the number of data points already assigned to any given cluster is comparable to the value of γ. Let u_c denote the number of evaluators assigned to a cluster c and let $\widetilde{u}_d$ denote the number of items assigned to a cluster d. The probability distribution determining the cluster assignment of evaluator j is given by:

$$P(c_j = c|\cdot) \propto u_c, \text{ if } c \leq L_1 \text{ is an existing cluster}$$
$$\propto \gamma, \text{ if } c > L_1 \text{ is a new cluster}$$

Similarly, the probability distribution determining the cluster assignment of item i is then given by:

$$P(d_i = d|\cdot) \propto \widetilde{u}_d, \text{ if } d \leq L_2 \text{ is an existing cluster}$$
$$\propto \gamma', \text{ if } d > L_2 \text{ is a new cluster}$$

3 Inference

In this section, we present an inference procedure for the proposed framework. Due to space constraints, we only discuss the derived conditional distributions for the most general Joint Confusion model.

Inference for Joint Confusion. The inference process involves estimating the conditional distribution over the set of hidden variables given all the observed variables. Exactly computing this posterior is intractable. Hence, we resort to approximate inference using Gibbs sampling where the conditional distribution is computed for each hidden variable based on the current assignments for all the other variables. In addition, we employ a variant of Gibbs sampling known as collapsed Gibbs sampling [14]. This enables faster convergence and mixing of the sampling chain by integrating out all the latent variables except for the cluster labels and item labels (c_j, z_i, d_i).

In the following expressions, $\boldsymbol{c}$ denotes the cluster assignments of all the evaluators, $\boldsymbol{d}$ denotes the cluster assignments of all the items, and $\boldsymbol{z}$ denotes all the item

labels. Similarly, we let $\boldsymbol{r}$ denote the evaluations, and $\boldsymbol{a}$ and $\boldsymbol{b}$ denote all the attributes of evaluators and items, respectively. Any variable superscripted with $-j$ and $-i$ indicates that evaluator j and item i are excluded from the counts under consideration.

The three hidden variables that we want to estimate in this model are $\boldsymbol{c}$, $\boldsymbol{d}$, and $\boldsymbol{z}$. In order to do so, we iterate over each of the variables associated with an individual evaluator or item and sample their values by holding the assignments of all the other latent variables intact. The conditional distribution for the cluster assignment c_j of evaluator j is given by:

$$P(c_j = c|\boldsymbol{c}^{-j}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{a}) \propto P(c_j = c|\boldsymbol{c}^{-j})$$
$$\times \prod_{items\ i\ labeled\ by\ j} P(r_{i,j}|\boldsymbol{r}^{-j}, \boldsymbol{c}, \boldsymbol{z}) \times \prod_{n=1}^{N} P(a_n^{(j)}|\boldsymbol{a}^{-j}, \boldsymbol{c})$$
$$= (u_c^{-j} + \epsilon_\alpha) \times \prod_{items\ i\ labeled\ by\ j} \left(\frac{v^{-j}_{z_i, r_{i,j}, c, d_i} + \wedge}{\sum_{w=1}^{L_1} (v^{-j}_{z_i, r_{i,j}, w, d_i} + \wedge)} \right)$$
$$\times \prod_{n=1}^{N} \left(\frac{x^{-j}_{a_n^{(j)}, c} + \epsilon_\beta}{\sum_{y \in\ values\ of\ a_n} (x^{-j}_{y,c} + \epsilon_\beta)} \right)$$

where u_c denotes the number of evaluators assigned to cluster c, $v_{z_i, r_{i,j}, c, d_i}$ denotes the number of times an item with label z_i is labeled as $r_{i,j}$ as per the confusion matrix for evaluator cluster c and the item cluster d_i, $x_{a_n^{(j)}, c}$ denotes the number of times the value of an attribute a_n is set to $a_n^{(j)}$ when the evaluator cluster is c.

The next hidden variable for which we need to determine the conditional distribution is d_i, which is the cluster assignment of item i and it is given by:

$$P(d_i = d|\boldsymbol{d}^{-i}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{b}) \propto P(d_i = d|\boldsymbol{d}^{-i})$$
$$\times \prod_{evaluators\ j\ labeling\ i} P(r_{i,j}|\boldsymbol{r}^{-i}, \boldsymbol{d}, \boldsymbol{z}^{-i}) \times \prod_{m=1}^{M} P(b_m^{(i)}|\boldsymbol{b}^{-i}, \boldsymbol{d})$$
$$= (\widetilde{u}_d^{-i} + \epsilon_{\alpha'}) \times \prod_{evaluators\ j\ labeling\ i} \left(\frac{v^{-i}_{z_i, r_{i,j}, c_j, d} + \wedge}{\sum_{w=1}^{L_2} (v^{-i}_{z_i, r_{i,j}, c_j, w} + \wedge)} \right)$$
$$\times \prod_{m=1}^{M} \left(\frac{\widetilde{x}^{-i}_{b_m^{(i)}, d} + \epsilon_{\beta'}}{\sum_{y \in\ values\ of\ b_m} (\widetilde{x}^{-i}_{y,d} + \epsilon_{\beta'})} \right)$$

where $\widetilde{u}_d$ denotes the number of items assigned to cluster d, $v_{z_i, r_{i,j}, c_j, d}$ denotes the number of times an item with label z_i is labeled as $r_{i,j}$ as per the confusion matrix for evaluator cluster c_j and the item cluster d, $\widetilde{x}_{b_m^{(i)}, d}$ denotes the number of times the value of an attribute b_m is set to $b_m^{(i)}$ when the item cluster is d. As before, the other symbols are defined in Table 1.

Last, we present the conditional distribution of z_i, the label of item i; this is given by:

$$P(z_i = z|\boldsymbol{z}^{-i}, \boldsymbol{r}, \boldsymbol{d}, \boldsymbol{c}) \propto P(z_i = z|\boldsymbol{z}^{-i}, \boldsymbol{d})$$
$$\times \prod_{evaluators\ j\ labeling\ i} P(r_{i,j}|\boldsymbol{r}^{-i}, \boldsymbol{d}, \boldsymbol{z}, \boldsymbol{c})$$
$$= (\bar{u}^{-i}_{d_i, z} + \epsilon_\rho) \times \prod_{evaluators\ j\ labeling\ i} \left(\frac{v^{-i}_{z, r_{i,j}, c_j, d_i} + \wedge}{\sum_{w=1}^{K} (v^{-i}_{w, r_{i,j}, c_j, d_i} + \wedge)} \right)$$

where $\bar{u}_{d_i, z}$ denotes the number of items belonging to cluster d_i that have true label z, $v_{z, r_{i,j}, c_j, d}$ denotes the number of times an item with label z is labeled as $r_{i,j}$ as per the confusion matrix for evaluator cluster c_j and item cluster d_i.

Estimating confusion matrices. Once we have inferred z_i, it is just a matter of computing the confusion matrices using inferred z_i and the already observed $r_{i,j}$. An entry in the confusion matrix for an evaluator j is computed using the expression:

$$\Theta^{(j)}_{s,t} = \frac{\sum_{i\ labeled\ by\ j} I(z_i = s) I(r_{i,j} = t)}{\sum_{i\ labeled\ by\ j} I(z_i = s)} \tag{3.1}$$

where $I(\cdot)$ is an indicator function. Intuitively, the expression above denotes the fraction of items for which the evaluator gave a label t when the actual label of the item was s. Note that this expression establishes the connection between the confusion matrix and the latent variable z_i that we estimated. Confusion matrices corresponding to evaluator and item clusters can be computed analogously.

4 Experimental Validation

Here, we discuss the experiments with our models on a variety of datasets. First, we focus on the quantitative analysis of the proposed framework. We establish that the proposed models actually capture the underlying dynamics of human evaluations. Then, we discuss various qualitative insights that we obtain by applying our models to several real-world datasets.

Dataset description. We experimented with several synthetic and real-world datasets to evaluate our models. Here, we present more details about each of the real-world datasets. Firstly, we analyze a sample of the data covering ***student examinations*** from two different courses on Coursera: Machine Learning and Algorithms. The answers given by several students for a subset of questions in each of these courses is recorded. Further, the set of all the correct answers for these questions constitute the ground-truth for item labels. Our second dataset comprises of ***peer grading*** activity on a subset of questions of the HCI course offered on Coursera. Each homework in this dataset was graded as either poor, basic, or good by a subset of student peer graders.

Dataset	# of Evaluators	# of Items	# of Decisions	Evaluator Attributes	Item Attributes
Student (Algo)	2,000	54	108,000	education, occupation, geography,	topic, number of words,
exams (ML)	2,000	53	106,000	gender, purpose of taking the course	type (conceptual/numerical)
Peer grading (PG)	5,000	6,224	19,208	education, occupation, geography, gender, purpose of taking the course	topic, week number, # of words, # of nouns, verbs, adjectives
Text labeling (Text)	152	4,000	11,400	gender, geography, age, self-reported confidence score	length, # of nouns, verbs, adjectives, TF-IDF vectors, # of named entities, topic specific words
Image labeling (Image)	101	450	3,915	gender, geography, age, self-reported confidence score	color of upper body, eye, tail, belly, shape of bill, wings

Table 6: Summary statistics of real-world data.

In addition, an expert rating is available for every homework and the set of all the expert ratings serve as the ground-truth for item labels. Each of these datasets contain various attributes pertaining to students, questions and answers(Table 6).

We also analyze the labeling activity on Amazon Mechanical Turk. We experiment with two different datasets — one where text documents need to be assigned appropriate topic labels and the other in which images need to be classified into different categories. In the ***text labeling*** task, each document needs to be assigned one of four possible topics: atheism, Christianity, baseball, and hockey. In the ***image labeling*** task, each image should be classified into one of the following categories: rusty blackbird, yellow-headed blackbird, brewer blackbird, and gray catbird. Several attributes of labelers, text and images are recorded in these datasets(Table 6).

Baselines. We compare our models against the following state of the art models which have been proposed to analyze the quality of judgments of individual evaluators: Single Confusion model (SC) [12], Dawid-Skene model (DS) [10], and Hybrid Confusion model (HC) [12]. Recall that these models are not set up to enable analysis based on evaluator or item attributes, but they serve as good candidates for benchmarking the performance of our algorithms in estimating confusion matrices, item labels, and predicting labels assigned by an evaluator.

Experimental setup. In most real-world settings where the goal is to analyze the quality of evaluators and also gain insights into their evaluations, we often do not have explicit access to the confusion matrices. Further, ground-truth labels are available only for a small subset of items. In our experiments, we simulate this setting by employing *weak supervsion*. The models are allowed access to true labels (in addition to the evaluator decisions) for only 15% of the items, while for the remaining 85% of the items, the models only have access to evaluator decisions (but not the true item labels). We run the inference process until the (approximate) convergence of log-likelihood. The hyperparameters are intialized to: $\epsilon_\rho = 0.01$ $\epsilon_\alpha = \epsilon_{\alpha'} = 0.2$, $\epsilon_{\beta'} = \epsilon_\beta = 0.1$, $\wedge = 1$, $\gamma = \gamma' = \frac{J \times I}{30}$. In addition, the number of clusters for evaluators and items were all determined using Bayesian Information Criterion (BIC) for the parametric versions of the models. On the other hand, the nonparametric variants automatically detect the number of clusters for both the evaluators and items.

Recovering true item labels. We evaluate the accuracy of each of the models on the task of recovering true item labels (Table 7 Left). We also experiment with logistic regression model (LR) as one of the baselines. In order to ensure a fair comparison between our models and a supervised approach such as logistic regression (LR), we use a randomly chosen subset of 15% of the data for training LR model and evaluate the model performance on the remaining 85% of the data. We use all the attributes of evaluators and items as independent variables in the prediction task. It can be seen that Joint Confusion and its nonparametric variant outperform all the other models. Overall, these models result in 8-15% improvement in predicting true labels of items.

Logistic regression (LR) and Single Confusion (SC) models turn out to be the weakest baselines. Further analysis revealed that the training data was insufficient for the LR model to make accurate predictions. In addition, the poor performance of SC can be attributed to its modeling assumptions which force all the evaluators to have a similar evaluation style. Dawid Skene (DS) and Hybrid Confusion (HC) perform on par with each other. However, our models consistently outperform all the baselines indicating that grouping evaluators and items based on their attribute values and associated evaluations is an effective way of recovering true labels of items.

Recovering confusion matrices. We also evaluate how accurately our models are able to recover confusion matrices corresponding to each of the evaluators. In order to achieve this, we use a metric called Mean

	Model	Accuracy - predicting item labels					MAE - estimating confusion matrices					Accuracy - predicting evaluator decisions				
		Algo	ML	PG	Text	Image	Algo	ML	PG	Text	Image	Algo	ML	PG	Text	Image
Baselines	Emp.						0.41	0.38	0.53	0.51	0.37					
	SC	0.55	0.56	0.57	0.56	0.54	0.43	0.46	0.42	0.48	0.41	0.52	0.58	0.53	0.51	0.56
	DS	0.59	0.60	0.62	0.65	0.61	0.34	0.33	0.35	0.36	0.32	0.61	0.61	0.64	0.60	0.64
	HC	0.61	0.60	0.64	0.65	0.63	0.31	0.30	0.30	0.29	0.28	0.62	0.64	0.66	0.59	0.63
	LR	0.53	0.57	0.54	0.54	0.55						0.55	0.53	0.51	0.56	0.57
Our Models	IC	0.65	0.64	0.64	0.65	0.67	0.32	0.34	0.34	0.29	0.29	0.67	0.72	0.69	0.64	0.66
	EC	0.66	0.65	0.66	0.65	0.69	0.28	0.32	0.28	0.24	0.27	0.68	0.72	0.70	0.66	0.69
	JC	0.67	**0.68**	0.69	0.69	0.71	0.22	0.25	0.26	**0.23**	0.25	0.68	0.74	0.70	**0.71**	**0.71**
	NJC	**0.70**	**0.68**	**0.71**	**0.70**	**0.72**	**0.21**	**0.22**	**0.24**	**0.23**	**0.23**	**0.69**	**0.75**	**0.73**	**0.71**	0.70
	Gain	14.8	13.3	10.9	7.7	14.3	11.3	17.2	10.6	20.3	11.1	32.3	26.7	20	20.7	17.9

Table 7: Experimental results of Predicting True Labels; Estimating Confusion Matrices; Predicting Evaluator Decisions; Columns correspond to different datasets (Table 6) and rows to different algorithms (Section 2). Gain denotes the pct. improvement of Non-parametric Confusion Model (NJC) over the Hybrid confusion (HC) Baseline.

Absolute Error (MAE) for assessing the quality of each estimated confusion matrix. Mathematically, this is expressed as: $MAE(\Theta^{(j)}, \hat{\Theta}^{(j)}) = \frac{1}{K^2}\sum_{s=1}^{K}\sum_{t=1}^{K}|\hat{\Theta}_{s,t}^{(j)} - \Theta_{s,t}^{(j)}|$ where $\Theta^{(j)}$ corresponds to the empirical estimate of the confusion matrix obtained from the dataset and $\hat{\Theta}^{(j)}$ represents the estimate of the same by a given model. The error of a model is then given by: $\frac{\sum_{j=1}^{J} MAE(\Theta^{(j)},\hat{\Theta}^{(j)})}{J}$. While the models Single Confusion (SC), Dawid-Skene (DS), Hybrid Confusion (HC), Evaluator Confusion (EC) associate a single confusion matrix with each evaluator, Item Confusion (IC) and Joint Confusion (JC) associate multiple such confusion matrices with any given evaluator. In order to account for this, we simply apply the MAE metric to every evaluator/item pair each time choosing an appropriate confusion matrix.

We present the results for various real-world datasets in Table 7 (Center). We also empirically estimate the confusion matrices from the subset of true labels (15%) which we used to initialize our models (Emp. in Table 7). The confusion matrices corresponding to this subset are computed by simple counting as we have access to both the true labels as well evaluator decisions (Equation 3.1). It can be seen that Joint Confusion and its nonparametric variant consistently outperform all the other baselines. In addition, it is interesting to note that the confusion matrices estimated by our models significantly outperform the empirical estimates which were accessible to the models.

Predicting labels given by evaluators. Our framework also allows for predicting the label of evaluator j on an item i. Though this is a natural side effect of modeling the process of evaluators' labeling, it turns out to be very useful in practice in settings such as behavioral targeting. The way we perform this prediction task is that we carry out latent variable inference on about 90% of the data assuming that we have access to the evaluator decisions for this chunk of the data. After the latent variable estimation on the 90% of the data, we now have a handle on the confusion matrices of all the evaluators and all the estimated item labels. Now, we can use these estimates to predict evaluator decisions on the residual 10% of the data and repeat this process resulting in a 10-fold cross validation. We obtain the predictions of the decisions on the residual 10% of the data in each pass by sampling variables $r_{i,j}$ from the appropriate confusion matrices and running this sampling process for about a few hundred iterations. The results of this task are presented in Table 7 (Right).

We observe that Joint Confusion and its nonparametric variant outperform all the other models. Item Confusion and Evaluator Confusion models almost perform similarly across all the datasets. As with all the other tasks, Single Confusion (SC) is the worst performing baselines clearly indicating that it is too restrictive to model all the decisions using a single confusion matrix for all the evaluators.

Qualitative analysis. All the models that we proposed so far produce explicit confusion matrices. The final step we need to discuss is the interpretation of the clusters. To interpret clusters, we represent each cluster by its *prominent attribute values*. To illustrate how we compute the prominent attribute values, let us consider an attribute *gender* and a value *female* in the student examinations dataset. Let $s^c_{G=F}$ denote the number of evaluators in the cluster c for whom the attribute gender has the value *female*. Further, let $s^c_{\cdot}$ denote the total number of students grouped into cluster c and let $s_{G=F}$ denote all the students in all the clusters for whom the attribute gender has the value *female*. The attribute value *gender = female* qualifies as a prominent attribute value if and only if $\frac{s^c_{G=F}}{s^c_{\cdot}} \geq \psi$ and $\frac{s^c_{G=F}}{s_{G=F}} \geq \omega$ where

ψ and ω correspond to support thresholds. Once we determine all the prominent attribute values for individual attributes, we also compute the above equation for combinations of these attributes (e.g. *gender = female and occupation = student*) and check if they can be added to the list of prominent attribute values. After this process, we obtain a list of prominent attribute values defining each cluster and we use them for describing high-level observations. We set the values of $\psi = (\frac{1.5}{\#\text{ of possible attribute values}})$ and ω to 0.80 throughout this analysis. We then relate these prominent attributes to the associated confusion matrices to understand the mistakes.

Analysis of items. We first use the Item Confusion model to obtain insights into the patterns of items that get evaluated in similar ways. In the text labeling dataset, lot of mistakes were made in tagging documents with a small number of sentences (<20). A lot of errors occur due to confusions between documents on atheism and Christianity, and between baseball and hockey (Figure 2(a)). In the peer grading dataset, we found that lengthy answers (>100 words) with lots of nouns (comprising >15% of all words) received more lenient grading; In the student examinations dataset, the probability of answering questions correctly in weeks 7, 8, 9 was much lower compared to all the other weeks (the probability for the ML class is 0.48). And, in the image labeling dataset, on images for which the color of the upper body and belly is gray, there were many mistakes in choosing between brewer blackbird, rusty blackbird, and gray catbird.

Analysis of evaluators. Next, we discuss some of the qualitative insights concerning evaluators which we obtain using the Evaluator Confusion model. In the text labeling dataset, we found that the cluster corresponding to females with low self-reported confidence scores is the cluster that results in the fewest mistakes compared to all the other clusters (Figure 2(b)). In the peer grading dataset, we found that male students who are currently pursuing undergraduate education are a lot stricter in their evaluations of peer assignments. In the student examinations dataset, we found that students enrolled in high-school and undergraduate studies have a lower probability of answering questions right (ML class probability = 0.23, Algorithms class probability = 0.31).

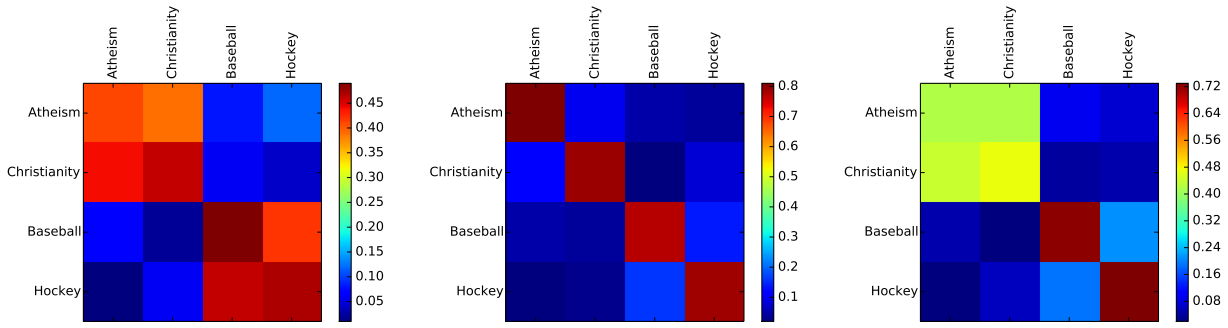
Analysis of evaluators and items. Finally, we use the Joint Confusion and its nonparametric variant to detect patterns that involve interplay between evaluators and items. In the text labeling dataset, male evaluators from Asia often confused documents on atheism and Christianity when there were fewer sentences of text (<20 sentences). On the other hand, they were able to very accurately tag documents belonging to baseball and hockey even with fewer sentences (Figure 2(c)). In the peer grading dataset, we found that students with an education level of masters degree and beyond from the US were very accurate in grading lengthy answers (>100 words). In the image labeling dataset, male evaluators belonging to the older age group (age > 50) often made mistakes when distinguishing between items for which a bird's bill is cone-shaped and upper body color is gray.

5 Related Work

The growth of crowdwork applications has catalyzed interest in two related but distinct research questions concerning human evaluations: inferring true labels from multiple annotations [1]-[7] and estimating evaluator expertise [8]-[12].

There have been attempts at unifying the two research directions by modeling evaluators and items [9, 15, 16, 17]. [9] associates an expertise level with individual evaluators and a difficulty level with each of the items; the probability that an evaluator rates an item correctly is then modeled as a function of these two parameters. This results in a model that can recover both the true labels and an estimate of the evaluator's quality, but since this approach does not quantify the underlying confusion-matrix structure, there is no direct way to build on it for obtaining diagnotic insights into the mistakes. [15] and [16] generalize the approach of [9] to account for characteristics of items and evaluators, though also without modeling the structure of the confusion matrix. Conversely, the models proposed in [10] and [12] estimate confusion matrices, but only at the level of individual evaluators, and so cannot produce collective insights about larger groupings within the evaluator population. There have also been recent approaches using information such as self-reported evaluator confidence scores to estimate evaluation quality and infer true labels [18].

In addition to producing cluster-level estimates of confusion matrices, our work allows for data in which each item may have been seen by only a few evaluators, each evaluator may have only looked at a small number of items, and ground truth may be missing for many of the items. Much of the earlier work relies on stronger assumptions about the data. For instance, [17] and [8] assume that each item has been annotated by most of the evaluators under consideration, which may not be the case in many natural settings. Approaches that work with evaluator-item pairs (e.g. [9, 19]) generally assume enough data to model each evaluator, which again is often not the case. Our approach also scales smoothly with the addition of new data, whereas much



(a) (b) (c)

Figure 2: Text labeling. (a). Aggregate item analysis: Documents with number of sentences < 20. (b) Aggregate evaluator analysis: Females with low self reported confidence scores. (c) Aggregate evaluator-item analysis: Male evaluators from Asia and Documents with number of sentences < 20.

of the earlier work (e.g. [15, 12, 18]) requires re-estimation in the presence of new data.

6 Conclusion

In many settings, evaluators seek to assign labels to items that they encounter (images, pieces of text, answers to quiz questions), and their performance can then be assessed against the ground-truth label for each item. In contrast to earlier lines of research aimed at inferring true labels or aggregate error rates, we estimate full confusion matrices at the level of both individual evaluators and items, and clusters of these. Understanding mistakes at the level of classes, across groups of evaluators with similar behavior, can provide insights into the structure of these mistakes.

The framework we present thus suggests a number of directions for further work. In particular, by discovering clusters of common mistake patterns, we can potentially identify interventions that can improve the performance of the underlying application. For example, in settings where it is possible to route items to particular evaluators, we can use a model of confusion matrix structure to find evaluators who will be particularly effective based on characteristics of the items.

References

[1] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, Learning from crowds. *JMLR*, pp. 1297–1322, 2010.

[2] V. S. Sheng, F. Provost, and P. G. Ipeirotis, Get another label? improving data quality and data mining using multiple, noisy labelers. *KDD*, pp. 614–622, 2008.

[3] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. *EMNLP*, pp. 254–263, 2008.

[4] O. Dekel and O. Shamir, Vox populi: Collecting high-quality labels from a crowd. *COLT*, 2009.

[5] S. Ertekin, H. Hirsh, and C. Rudin, Learning to predict the wisdom of crowds. *Collective Intelligence*, 2011.

[6] D. R. Karger, S. Oh, and D. Shah, Iterative learning for reliable crowdsourcing systems. *NIPS*, pp. 1953–1961, 2011.

[7] E. Kamar, S. Hacker, and E. Horvitz, Combining human and machine intelligence in large-scale crowdsourcing. *AAMAS*, pp. 467–474, 2012.

[8] M. Joglekar, H. Garcia-Molina, and A. Parameswaran, Evaluating the crowd with confidence. *KDD*, pp. 686–694, 2013.

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

[10] A. P. Dawid and A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, no. 1, pp. 20–28, 1979.

[11] H. Li, B. Yu, and D. Zhou, Error rate analysis of labeling by crowdsourcing. *ICML*, 2013.

[12] C. Liu and Y.-M. Wang, Truelabel + confusions: A spectrum of probabilistic models in analyzing multiple ratings. *ICML*, 2012.

[13] C. E. Antoniak, Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, vol. 2, no. 6, 1974.

[14] T. L. Griffiths and M. Steyvers, Finding scientific topics. *PNAS*, pp. 5228–5235, 2004.

[15] P. Welinder, S. Branson, S. Belongie, and P. Perona, The multidimensional wisdom of crowds. *NIPS*, pp. 2424–2432, 2010.

[16] P. Welinder and P. Perona, Online crowdsourcing: rating annotators and obtaining cost-effective labels. *Workshop on Advancing Computer Vision with Humans in the Loop, CVPR*, 2010.

[17] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, Supervised learning from multiple experts: Whom to trust when everyone lies a bit. *ICML*, 2009.

[18] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima, Accurate integration of crowdsourced labels using workers' self-reported confidence scores. *IJCAI*, 2013.

[19] D. Zhou, J. C. Platt, S. Basu, and Y. Mao, Learning from the wisdom of crowds by minimax entropy. *NIPS*, pp. 2204–2212, 2012.