Lecture 6:
CS 5306 / INFO 5306:
Crowdsourcing and
Human Computation

Two Qualitatively Different Types of Tasks

- Objective
  - There is a correct “ground truth”
  - Example: Patient has cancer
- Cultural / Subjective
  - Different people may legitimately differ on answer
  - Example: Is an image pornographic?

Aggregating Crowdsourced Creations

- Ask different people to do the identical task, resolve difference in answers
  Why:
  - Subjective differences
  - Fallibility
  - Malevolence
- Examples:
  - ESP Game: multiple random pairs need to agree
  - MTurk: Take majority vote

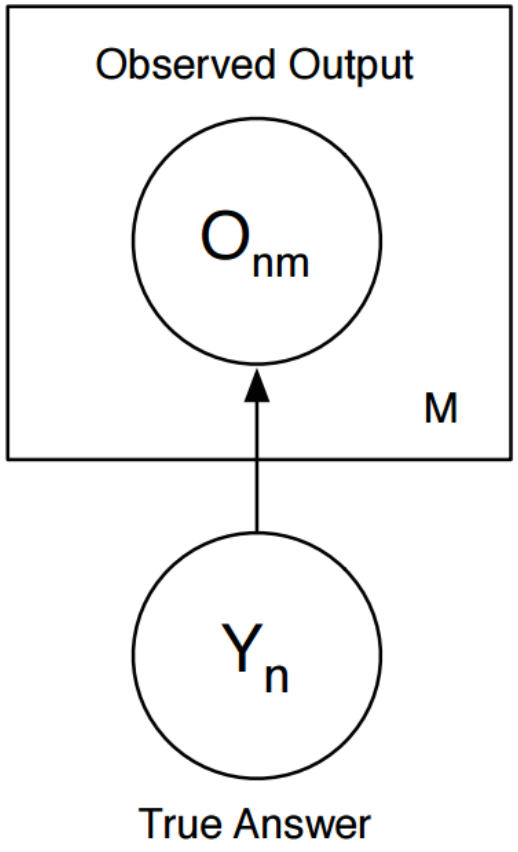
$$
\begin{array}{c}
 \\
 \\
\text{worker}
\end{array}
\begin{array}{c}
\text{computational task} \\
\begin{array}{cccc} 1 & 2 & \cdots & N \end{array} \\
\begin{array}{c} 1 \\ 2 \\ \vdots \\ M \end{array}
\begin{bmatrix}
O_{11} & O_{12} & \cdots & O_{1N} \\
O_{21} & O_{22} & \cdots & O_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
O_{M1} & O_{M2} & \cdots & O_{MN}
\end{bmatrix}
\end{array}
$$

$$
\begin{aligned}
Y_n &= \operatorname*{argmax}_j P(Y_n = j|O) \\
&= \operatorname*{argmax}_j \frac{\prod_{m=1}^{M} P(O_{n,m} = o_{n,m}|Y_n = j)P(Y_n = j)}{P(O)} \\
&\propto \operatorname*{argmax}_j \prod_{m=1}^{M} P(O_{n,m} = o_{n,m}|Y_n = j) \\
&\propto \operatorname*{argmax}_j (1-\epsilon)^{\sum_{m=1}^{M} \mathbf{1}(o_{n,m}=j)} \cdot \epsilon^{\sum_{m=1}^{M} \mathbf{1}(o_{n,m}\neq j)}
\end{aligned}
$$

Majority Vote

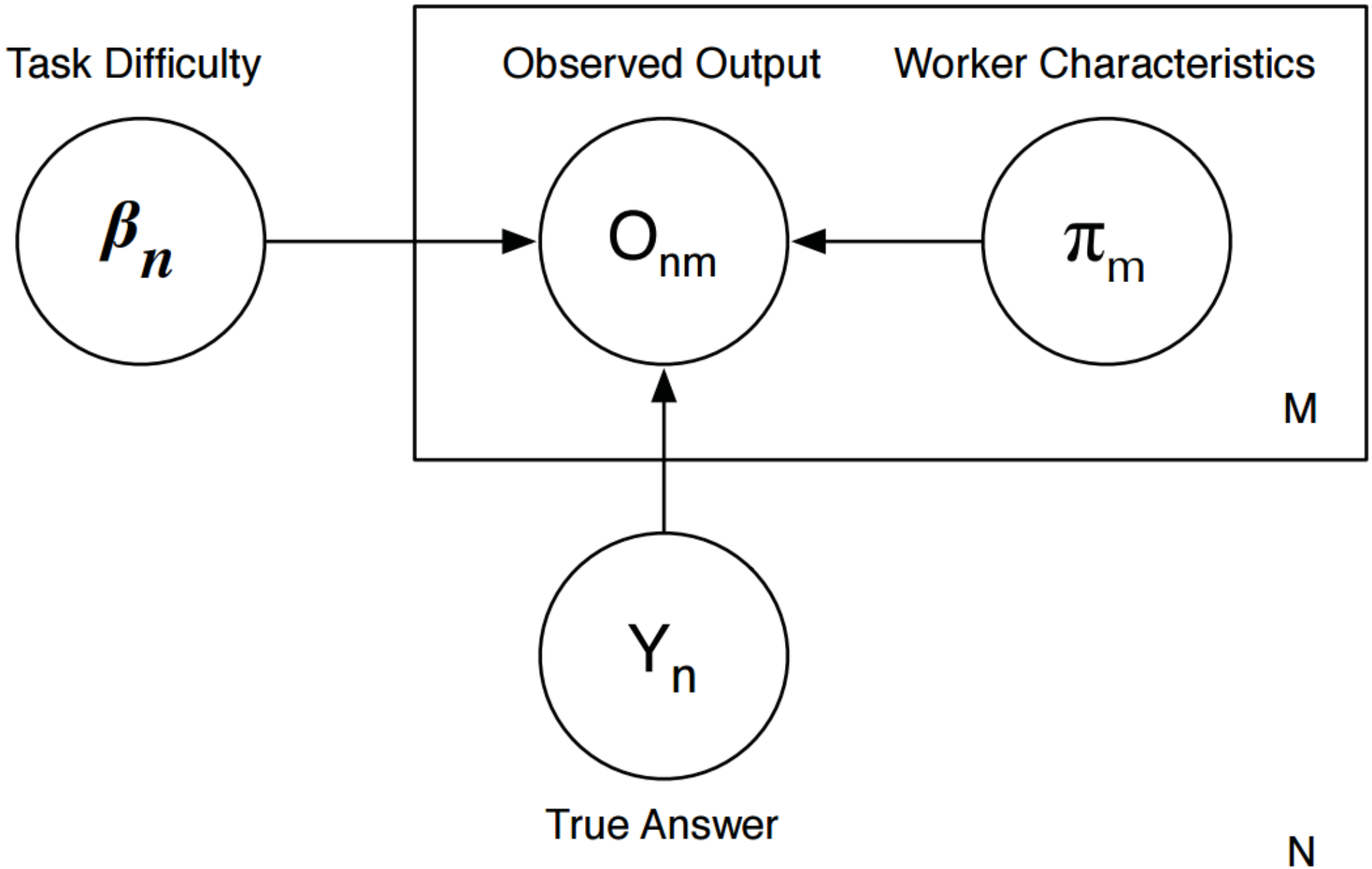
$$
\begin{aligned}
Y_n &= \operatorname*{argmax}_j P(Y_n = j|O) \\
&= \operatorname*{argmax}_j \frac{\prod_{m=1}^{M} P(O_{n,m} = o_{n,m}|Y_n = j)P(Y_n = j)}{P(O)} \\
&\propto \operatorname*{argmax}_j \prod_{m=1}^{M} P(O_{n,m} = o_{n,m}|Y_n = j) \\
&\propto \operatorname*{argmax}_j (1-\epsilon)^{\sum_{m=1}^{M} \mathbf{1}(o_{n,m}=j)} \cdot \epsilon^{\sum_{m=1}^{M} \mathbf{1}(o_{n,m}\neq j)}
\end{aligned}
$$

Graphical Model Representation



Probabilistic Models for Aggregation

- Can incorporate:
  - Different problem difficulties (e.g., more or fewer votes needed)
  - Differential voter ability (e.g., can weight votes)
  - Differential worker competence (e.g., different people are better on different questions)
- When not known, they are “latent variables” that are inferred from the data
- Common approach: Expectation Maximization


Task Difficulty
Observed Output
Worker Characteristics
β_n
O_{nm}
π_m
M
Y_n
True Answer
N

TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings

Chao Liu[†] CRAIGLIU@TENCENT.COM
Tencent Inc, 38 Haidian St, Beijing, 100080, P. R. China

Yi-Min Wang YMWANG@MICROSOFT.COM
Microsoft Research, 1 Microsoft Way, Redmond, WA 98052, USA

Abstract

This paper revisits the problem of analyzing multiple ratings given by different judges. Different from previous work that focuses on distilling the true labels from noisy crowdsourcing ratings, we emphasize gaining diagnostic insights into our in-house well-trained judges. We generalize the well-known DawidSkene model (Dawid & Skene, 1979) to a spectrum of probabilistic models under the same "TrueLabel + Confusion" paradigm, and show that our proposed hierarchical Bayesian model, called HybridConfusion, consistently outperforms DawidSkene on both synthetic and real-world data sets.

1. Motivation

Recent advent of online crowdsourcing services (*e.g.*, Amazon's Mechanical Turk) excites the machine learning community by making large amount of labeled data practical. Because of the low cost, crowdsourcing labels are usually given by anonymous lowly-paid non-experts, which sparks recent interest in recovering the true labels from noisy (or even malicious) labels (Whitehill et al., 2009; Welinder et al., 2010; Welinder & Perona, 2010; Raykar et al., 2009). In this paper, we study the same problem of analyzing multiple ratings, but in quite a different setting.

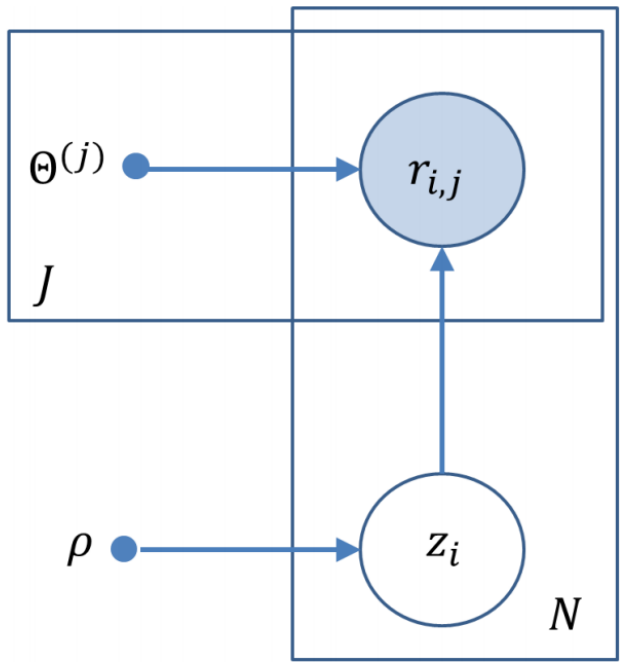
We are in a major Web search engine company, and train search rankers using human ratings on the relevance of tens of millions of (query, URL) pairs. As it is too risky to bet the search engine on crowdsourcing ratings, we have to carefully recruit human judges, rigorously train them, and continually monitor their quality during the work. Since these judges are well-trained and the rating task is considerably hard, the cost of each label becomes so expensive that even two ratings per (query, URL) pair are economically infeasible: note that we have millions of pairs to rate and the number keeps increasing. Instead, we hope that a human judge would function satisfactorily once qualified, and each (query, URL) pair is only rated by one judge.

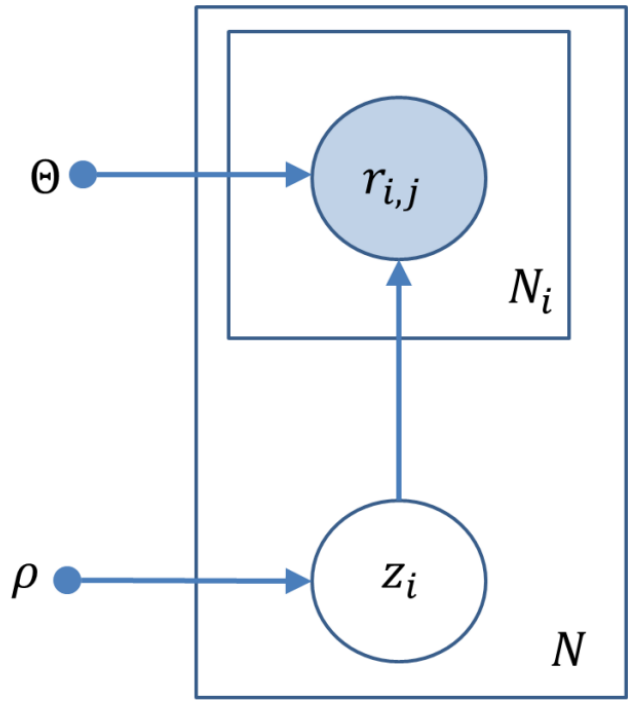
A key component in controlling the judge quality is to blend a small set of "monitoring" (query, URL) pairs into judges' regular work without their knowledge. This set of (query, URL) pairs are rated by all judges under monitoring. By analyzing the multiple ratings on (query, URL) pairs in this monitoring set, we hope to correctly score the quality of each judge, and more importantly, to gain insights into what confusions each judge makes so that we could plan targeted tutoring and revisions to the rating guidelines. Therefore, different from previous work that focuses on recovering the true labels from low-cost noisy labels, we are more interested in diagnostic information about judge confusions. For this reason, this paper emphasizes on probabilistic models that use a confusion matrix to quantify the competency of each judge.

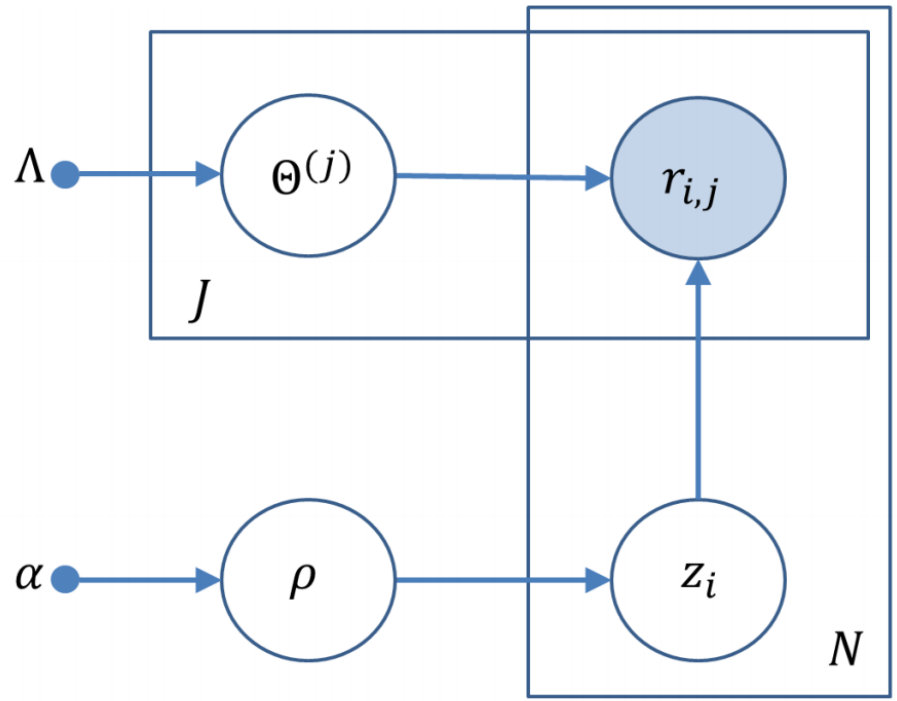
The DawidSkene model (Dawid & Skene, 1979) is a good candidate for this purpose. It pioneers the "TrueLabel + Confusion" paradigm: each item has a true label, and the rating each judge assigns to it is the true label obfuscated through the judge's confusion matrix. Suppose the rating is on a K-level scale, a confusion matrix is a $K \times K$ matrix

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[†] This work was done when the first author was employed by Microsoft Research at Redmond.


$\Theta^{(j)}$
$r_{i,j}$
J
ρ
z_i
N


Θ
$r_{i,j}$
N_i
ρ
z_i
N


Λ
$\Theta^{(j)}$
$r_{i,j}$
J
α
ρ
z_i
N

Latent Confusion Analysis by Normalized Gamma Construction

Issei Sato SATO@R.DL.ITC.U-TOKYO.AC.JP
The University of Tokyo

Hisashi Kashima KASHIMA@I.KYOTO-U.AC.JP
Kyoto University

Hiroshi Nakagawa N3@DL.ITC.U-TOKYO.AC.JP
The University of Tokyo

Abstract

We developed a flexible framework for modeling the annotation and judgment processes of humans, which we called "normalized gamma construction of a confusion matrix." This framework enabled us to model three properties: (1) the abilities of humans, (2) a confusion matrix with labeling, and (3) the difficulty with which items are correctly annotated. We also provided the concept of "latent confusion analysis (LCA)," whose main purpose was to analyze the principal confusions behind human annotations and judgments. It is assumed in LCA that confusion matrices are shared between persons, which we called "latent confusions", in tribute to the "latent topics" of topic modeling. We aim at summarizing the workers' confusion matrices with the small number of latent principal confusion matrices because many personal confusion matrices is difficult to analyze. We used LCA to analyze latent confusions regarding the effects of radioactivity on fish and shellfish following the Fukushima Daiichi nuclear disaster in 2011.

1. Introduction

An important theme in collective intelligence is modeling the annotation and judgment processes of humans. We focus on modeling a confusion matrix with labeling. Extracting a confusion matrix is useful for not just obtaining better (closer to the ground truth) aggregation of labels but also obtaining diagnostic information on human annotation and judgments.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Dawid and Skene (1979) proposed a probabilistic generative model for subjective labeling. Their model can estimate individual confusion matrices even when the true label is not available. Each worker in this model has a confusion matrix in which if an item (e.g., an image) has true label u, worker j can assign another label l with probability $\pi_{u,l}^{(j)}$. Smyth et al. (1994) applied the Dawid and Skene (DS) model to the image labeling problem. Snow et al. (2008) applied the DS model to the analysis of opinions in natural language processing. Liu and Wang (2012) applied the DS model to judge the quality of (query, URL) pairs.

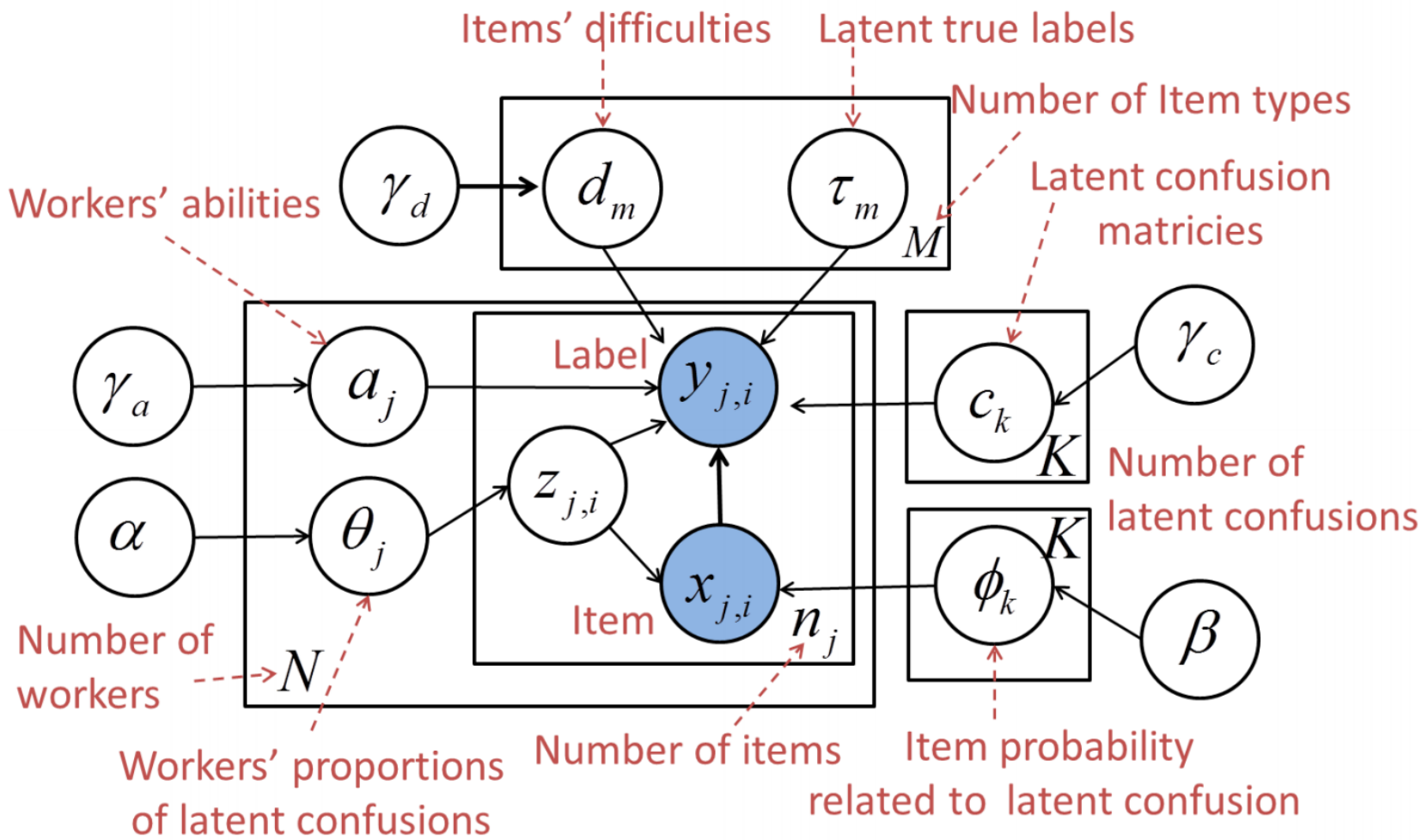
Whitehill et al. (2009) proposed the Generative model of Labels, Abilities, and Difficulties (GLAD), which simultaneously estimated the expertise of each worker and the difficulty of each task. It is beneficial to use GLAD, unlike the DS model, in that it models the difficulty with which items are correctly annotated. However, it suffers from a critical issue that when we apply GLAD to a task with multiple labels, the confusion matrix of a worker cannot be constructed (see Sec.3.2 for the details).

Contributions: This paper makes three contributions.

(1) We propose a normalized gamma construction (NGC) of a confusion matrix to model the annotation and judgment process of humans. This framework easily enables us to model a confusion matrix with labeling in a multi-label setting like the DS model and to take into account a task's difficulty like that with GLAD.

(2) We provide a novel concept in data science, *latent confusion analysis (LCA)*, which was developed with the NGC framework and latent Dirichlet enhanced modeling. The main aim of LCA is to extract latent (principal) confusions behind the annotation and judgment processes of humans. LCA summarizes the workers' confusion matrices with the small number of latent principal confusion matrices because many personal confusion matrices is difficult to analyze.

(3) The proposed learning algorithm was based on the


Items' difficulties
Latent true labels
Number of Item types
Latent confusion matricies
Workers' abilities
γ_d
d_m
τ_m
M
γ_a
a_j
Label
$y_{j,i}$
c_k
γ_c
K
Number of latent confusions
$z_{j,i}$
α
θ_j
$x_{j,i}$
ϕ_k
K
β
Item
n_j
Number of workers
N
Workers' proportions of latent confusions
Number of items
Item probability related to latent confusion

Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan
Machine Perception Laboratory
University of California, San Diego
La Jolla, CA, USA
{ jake, paul, ting, jbergsma, movellan }@mplab.ucsd.edu

Abstract

Modern machine learning-based approaches to computer vision require very large databases of hand labeled images. Some contemporary vision systems already require on the order of millions of images for training (e.g., Omron face detector [9]). New Internet-based services allow for a large number of labelers to collaborate around the world at very low cost. However, using these services brings interesting theoretical and practical challenges: (1) The labelers may have wide ranging levels of expertise which are unknown a priori, and in some cases may be adversarial; (2) images may vary in their level of difficulty; and (3) multiple labels for the same image must be combined to provide an estimate of the actual label of the image. Probabilistic approaches provide a principled way to approach these problems. In this paper we present a probabilistic model and use it to simultaneously infer the label of each image, the expertise of each labeler, and the difficulty of each image. On both simulated and real data, we demonstrate that the model outperforms the commonly used "Majority Vote" heuristic for inferring image labels, and is robust to both noisy and adversarial labelers.

1 Introduction

In recent years machine learning-based approaches to computer vision have helped to greatly accelerate progress in the field. However, it is now becoming clear that many practical applications require very large databases of hand labeled images. The labeling of very large datasets is becoming a bottleneck for progress. One approach to address this incoming problem is to make use of the vast human resources on the Internet. Indeed, projects like the ESP game [17], the Listen game[16], Soylent Grid [15], and reCAPTCHA [18] have revealed the possibility of harnessing human resources to solve difficult machine learning problems. While these approaches use clever schemes to obtain data from humans for free, a more direct approach is to hire labelers online. Recent Web tools such as Amazon's *Mechanical Turk* [1] provide ideal solutions for high-speed, low cost labeling of massive databases.

Due to the distributed and anonymous nature of these tools, interesting theoretical and practical challenges arise. For example, principled methods are needed to combine the labels from multiple experts and to estimate the certainty of the current labels. Which image should be labeled (or relabeled) next must also be decided – it may be prudent, for example, to collect many labels for each image in order to increase one's confidence in that image's label. However, if an image is easy and the labelers of that image are reliable, a few labels may be sufficient and valuable resources may be used to label other images. In practice, combining the labels of multiple coders is a challenging process due to the fact that: (1) The labelers may have wide ranging levels of expertise which are

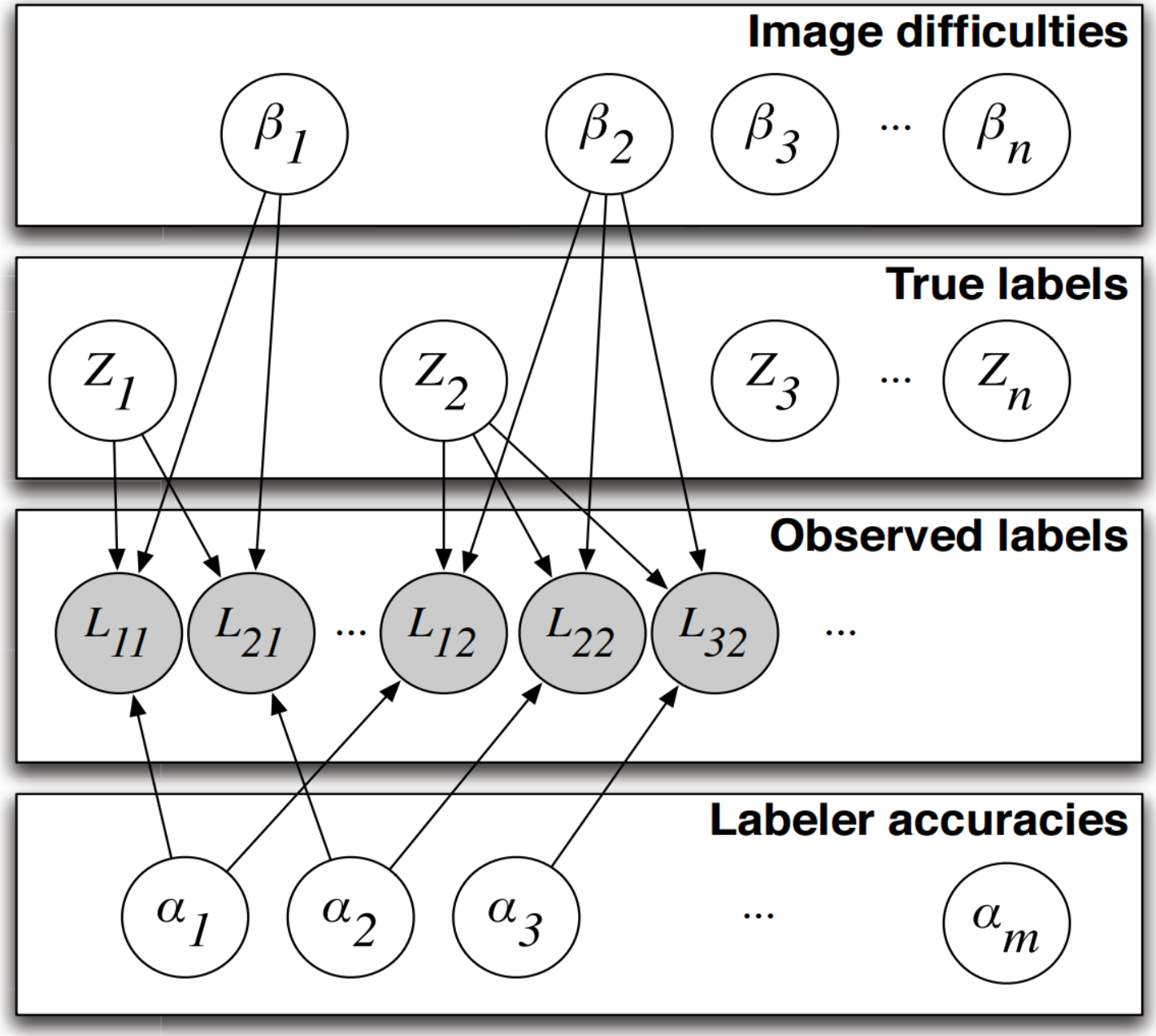


Image difficulties
β_1
β_2
β_3
...
β_n
True labels
Z_1
Z_2
Z_3
...
Z_n
Observed labels
L_{11}
L_{21}
...
L_{12}
L_{22}
L_{32}
...
Labeler accuracies
α_1
α_2
α_3
...
α_m

The Multidimensional Wisdom of Crowds

Peter Welinder[1] **Steve Branson**[2] **Serge Belongie**[2] **Pietro Perona**[1]
[1] California Institute of Technology, [2] University of California, San Diego
{welinder,perona}@caltech.edu {sbranson,sjb}@cs.ucsd.edu

Abstract

Distributing labeling tasks among hundreds or thousands of annotators is an increasingly important method for annotating large datasets. We present a method for estimating the underlying value (e.g. the class) of each image from (noisy) annotations provided by multiple annotators. Our method is based on a model of the image formation and annotation process. Each image has different characteristics that are represented in an abstract Euclidean space. Each annotator is modeled as a multidimensional entity with variables representing competence, expertise and bias. This allows the model to discover and represent groups of annotators that have different sets of skills and knowledge, as well as groups of images that differ qualitatively. We find that our model predicts ground truth labels on both synthetic and real data more accurately than state of the art methods. Experiments also show that our model, starting from a set of binary labels, may discover rich information, such as different "schools of thought" amongst the annotators, and can group together images belonging to separate categories.

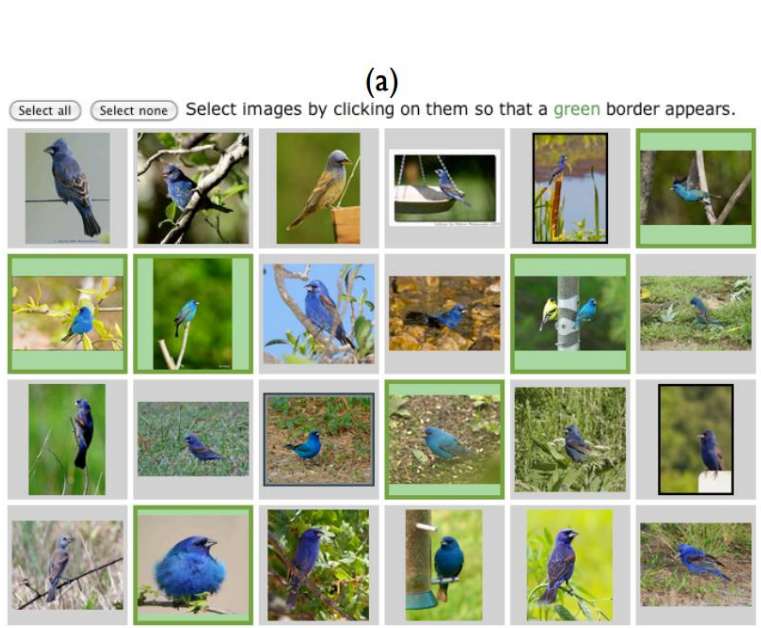
1 Introduction

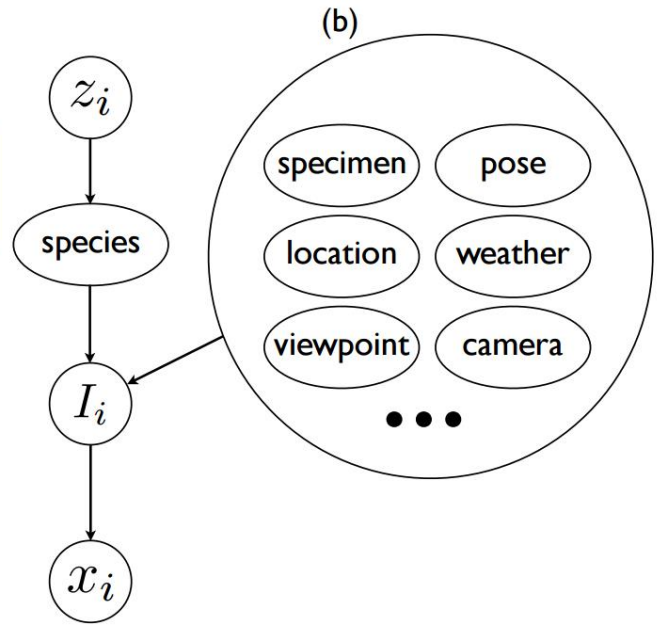
Producing large-scale training, validation and test sets is vital for many applications. Most often this job has to be carried out "by hand" and thus it is delicate, expensive, and tedious. Services such as Amazon Mechanical Turk (MTurk) have made it easy to distribute simple labeling tasks to hundreds of workers. Such "crowdsourcing" is increasingly popular and has been used to annotate large datasets in, for example, Computer Vision [8] and Natural Language Processing [7]. As some annotators are unreliable, the common wisdom is to collect multiple labels per exemplar and rely on "majority voting" to determine the correct label. We propose a model for the annotation process with the goal of obtaining more reliable labels with as few annotators as possible.

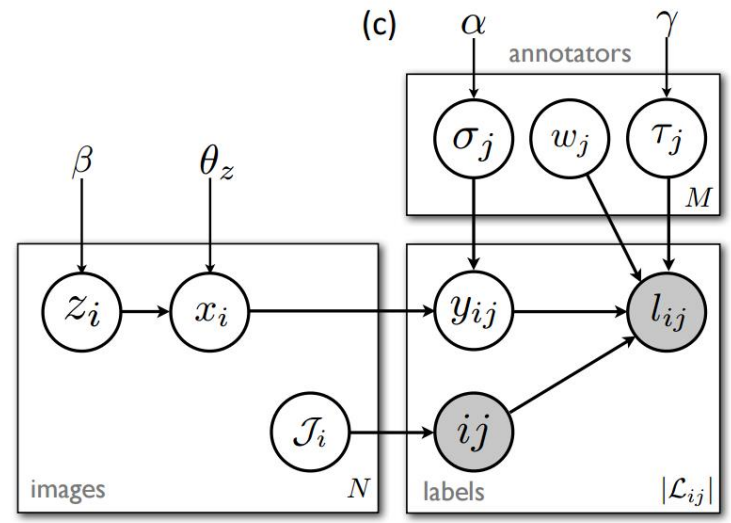
It has been observed that some annotators are more skilled and consistent in their labels than others. We postulate that the ability of annotators is multidimensional; that is, an annotator may be good at some aspects of a task but worse at others. Annotators may also attach different costs to different kinds of errors, resulting in different biases for the annotations. Furthermore, different pieces of data may be easier or more difficult to label. All of these factors contribute to a "noisy" annotation process resulting in inconsistent labels. Although approaches for modeling certain aspects of the annotation process have been proposed in the past [1, 5, 6, 9, 13, 4, 12], no attempt has been made to blend all characteristics of the process into a single unified model.

This paper has two main contributions: (1) we improve on current state-of-the-art methods for crowdsourcing by introducing a more comprehensive and accurate model of the human annotation process, and (2) we provide insight into the human annotation process by learning a richer representation that distinguishes amongst the different sources of annotator error. Understanding the annotation process can be important toward quantifying the extent to which datasets constructed from human data are "ground truth".

We propose a generative Bayesian model for the annotation process. We describe an inference algorithm to estimate the properties of the data being labeled and the annotators labeling them. We show on synthetic and real data that the model can be used to estimate data difficulty and annotator


(a)
Select all
Select none
Select images by clicking on them so that a green border appears.


(b)
z_i
species
I_i
x_i
specimen
pose
location
weather
viewpoint
camera


(c)
α
γ
annotators
σ_j
w_j
τ_j
M
β
θ_z
z_i
x_i
y_{ij}
l_{ij}
$\mathcal{J}_i$
ij
images
N
labels
$|\mathcal{L}_{ij}|$

A “Catalog” Characterization of Tasks

- Classification
  - Easiest
- Ranking
  - Chess
- Clustering
  - How do you combine different people’s categories?
- Structured Outputs
  - Language

Each may require different (and increasingly more complicated) forms of aggregation

Readings for Next Time

- Tuesday, February 23:
  Infotopia, Chapter 1