Outline

Motivation: Background

- Artificial intelligence are relying more and more on large-scale training datasets.
- Expert labeling are expensive and time-consuming.

Motivation: Crowdsourcing Solution

- Using multiple web workers to label each item.
- Recovery the ground truth from the noisy data.

Table: Different workers may give inconsistent labels to a same item.

	worker a	worker b	worker c	worker d
item 1	1	**2**	1	1
item 2	**2**	**2**	1	**2**
item 3	1	1	**2**	1
item 4	1	1	1	**2**

Outline

Majority Voting

Assumption

For every worker, the ground truth label is always the most common to be given, and the labels for each item are given independently.

$$\mathbb{P}(Y_m = d) = \frac{\sum_{(n,m)\in \mathbf{I}} \delta_{W_{nm},d}}{\sum_{d,(n,m)\in \mathbf{I}} \delta_{W_{nm},d}}, \forall m, \quad (1)$$

Dawid-Skene estimator

Assumption

The performance of a worker is consistent across different items, and his or her behavior can be measured by a confusion matrix.

Table: An example of binary classification confusion matrix.

	label A	label B
label A	1	0
label B	1/3	2/3

Dawid-Skene estimator

Assumption

The performance of a worker is consistent across different items, and his or her behavior can be measured by a confusion matrix.

$$\mathbb{P}(\mathbf{W}|\mathbf{q},\mathbf{p}) = \prod_m \left(\sum_d q_d \prod_{n,l} {p_{ndl}}^{\delta_{W_{nm},l}} \right), \qquad (2)$$

p_{ndl}: the probability that worker n labels an item as l when it's ground truth label is d.

Outline

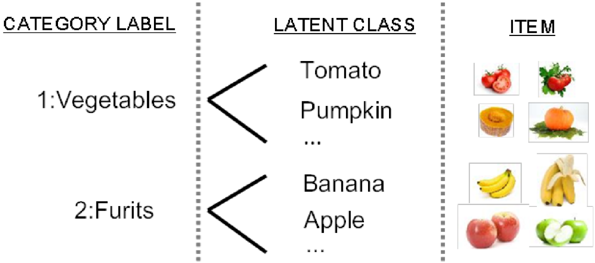
Latent Class Assumptions

- **Assumption I.** *Each item belongs to one specific latent class only.*
- **Assumption II.** *Items in the same latent class should be classified into the same category.*

Latent Class Assumptions

- Each item belongs to one specific latent class only.
- Items in the same latent class in the same category.

Figure: Latent classes of a binary classification task to classify fruit and vegetable.



Latent Class Confusion Matrix

Extend the original confusion matrix. An entry p_{nkl} in this matrix represents the probability that worker n gives label l to an item which belongs to latent class k.

Table: An example of latent class confusion matrix.

	tomato	pumpkin	cuke	apple
Fruit	2/3	0	1/3	1
Vegetable	1/3	1	2/3	0

Chinese Restaurant Process

- The number of latent classes K is unknown in advance.
- we choose a nonparametric prior for the latent class confusion matrix to learn the number of classes.

Z_i is the latent class assignment of item i, with the CRP prior, the probability

$$\begin{aligned} \text{Old Class}: \quad P(Z_i = k) &\propto n_k + \alpha, \forall k = 1 \cdots K, & (3) \\ \text{New Class}: P(Z_i = K+1) &\propto \alpha. & (4) \end{aligned}$$

Nonparametric Dawid-Skene Model

Full Bayesian Model:

$$\begin{aligned} \textit{Assign} &: \quad \mathbf{Z}|\alpha_c \sim \mathrm{CRP}(\alpha_c), && (5)\\ \textit{Entries} &: \quad \mathbf{p_{nk}}|\alpha_d \sim \mathrm{Dirichlet}(\alpha_d),\ \forall n, k, && (6)\\ \textit{Obervation} &: \quad W_{nm}|\mathbf{Z}, \mathbf{p}_{n\cdot} \sim \mathrm{Multinomial}(\mathbf{A}_{nm}),\ \forall n, m, && (7) \end{aligned}$$

Here $\mathbf{A}_{nm} = \{A_{nm1}, \cdots, A_{nmD}\}, A_{nmd} = \prod_{k=1}^{K} {p_{nkd}}^{\delta_{Z_m,k}}$

Inference

We use Gibbs Sampling to infer the hidden variables.
Conditional Distributions: confusion matrix parameter:

$$\mathbf{p}_{nk}|\mathbf{Z},\mathbf{W} \sim \text{Dirichlet}(\mathbf{p}_{nk}|\mathbf{B}_{nk}), \forall n, k, \tag{8}$$

where $B_{nkd} = \sum_{m=1}^{M} \delta_{W_{nm},d}\delta_{Z_m,k} + \alpha_d/D$.
hidden variables, when $k \leq K$,

$$P(Z_m = k|\mathbf{Z}_{-m},\mathbf{p},\mathbf{W}) \propto n_k \prod_{n=1}^{N}\prod_{d=1}^{D} {p_{nkd}}^{\delta_{W_{nm},d}}, \tag{9}$$

When generating a new class,

$$P(Z_m = k_{new}|\mathbf{Z}_{-m},\mathbf{p},\mathbf{W}) \propto \alpha_c \prod_{n=1}^{N} \frac{\prod_{d=1}^{D}\Gamma(\delta_{W_{nm},d} + \alpha_d/D)}{\Gamma(1+\alpha_d)}. \tag{10}$$

Latent Class Minimax Entropy Estimator

We also extend the minimax entropy estimator.
Idea: change the category constraints.
New objective:

$$\min_{\mathbf{Z}} \max_{\mathbf{p},\boldsymbol{\tau},\boldsymbol{\sigma}} -\sum_{n,m,d} p_{nmd} \log p_{nmd} - \sum_{m,d} \frac{\alpha_m \tau_{md}^2}{2} - \sum_{n,m,d} \frac{\beta_n \sigma_{ndk}^2}{2}$$

$$s.t. \sum_{n} \left(p_{nmd} - \delta_{W_{nm},d}\right) = \tau_{md}, \forall m, d,$$

$$\sum_{m} \left(p_{nmd} - \delta_{W_{nm},d}\right)\delta_{Z_{m},k} = \sigma_{ndk}, \forall n, k, \quad \sum_{d} p_{nmd} = 1, \forall n, m. \tag{11}$$

Category Recovery

- Regard items in a same class as one imaginary item, here we call it a *hyper-item*.
- We use a generalized Dawid-Skene estimator with hyper-items to estimate the category assignments.

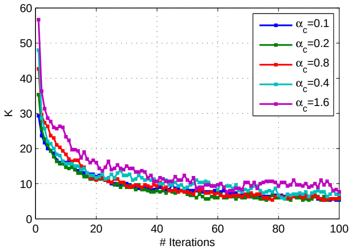
$$\mathbb{P}(\mathbf{W}|\mathbf{q},\mathbf{p}) = \prod_k \left(\sum_d q_d \prod_{n,l} {p_{ndl}}^{n_{nkd}} \right), \qquad (12)$$

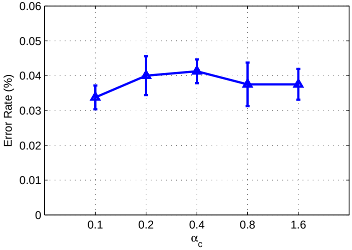
where $n_{nkd} = \sum_m \delta_{W_{nm},d}\delta_{Z_m,k}$ is the count of labels that worker n gives to hyper-item k.

Outline

Synthetic Dataset

- 4 latent classes, 40 items' parameters for each latent class.
- 2 types of workers, 20 workers of each type.



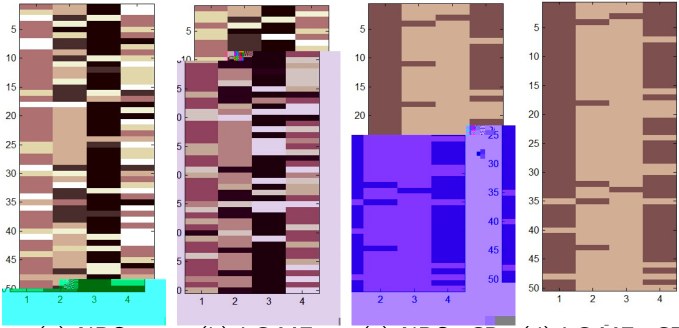


(a) (b)

Figure: (a) shows the numbers of latent classes found by NDS with different color. (b) shows the average category recovery error rates.

Real Flower Dataset

- 4 flower species, 50 pics for each. 2 species for each category.
- 36 workers, 2366 labels in total.



(a) NDS (b) LC-ME (c) NDS+CR (d) LC-ME+CR

Figure: entry corresponding to image, column corresponding to flower species. (a)(b): color denotes latent class. (c)(d): denotes category.

Real Flower Dataset

- 4 flower species, 50 pics for each. 2 species for each category.
- 36 workers, 2366 labels in total.

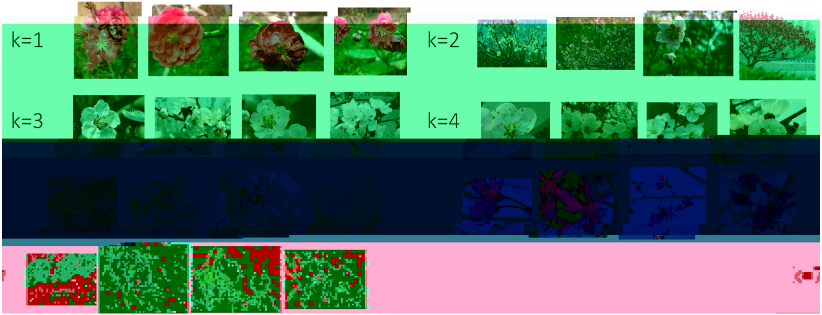


Figure: Representative pictures for different latent classes.(best viewed in color).

Category Recovery

Table: Performance of several models on flowers dataset. the average error rate of 10 trials, together with standard deviation, are presented.

#	20	25	30	35
MV	0.1998 ± 0.0506	0.2383 ± 0.0216	0.2153 ± 0.0189	0.2170 ± 0.0096
DS	0.1590 ± 0.0538	0.1555 ± 0.0315	0.1310 ± 0.0213	0.1300 ± 0.0041
NDS	0.1595 ± 0.0737	0.1605 ± 0.0434	0.1330 ± 0.0371	0.1475 ± 0.0354
ME	0.1535 ± 0.0695	0.1470 ± 0.0339	0.1315 ± 0.0200	0.1335 ± 0.0078
LC-ME	**0.1415 ± 0.0382**	**0.1430 ± 0.0286**	**0.1215 ± 0.0133**	**0.1190 ± 0.0168**

The first line means the number of workers we used on estimation tasks.

Outline

Outline

Selected Reference

Dawid, A.P., and Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Applied statistics (1979), 20–28.

Dempster, A.P., Laird, N.M., Rubin, D.B., etal.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal statistical Society 39, 1 (1977), 1–38.

Li, H., Yu, B., and Zhou, D.: Error rate analysis of labeling by crowdsourcing.

Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., and Moy, L.: Learning from crowds. The Journal of Machine Learning Research 11 (2010), 1297–1322.

Sheshadri, A., and Lease, M.: Square: A benchmark for research on computing crowd consensus. In First AAAI Conference on Human Computation and Crowdsourcing (2013).

Tian, Y., and Zhu, J.: Learning from crowds in the presence of schools of thought. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (2012), ACM, 226–234.

Welinder, P., Branson, S., Belongie, S., and Perona, P.: The multidimensional wisdom of crowds. In Neural Information Processing Systems (2010), vol.10, 2424–2432.

Zhou, D., Platt, J.C., Basu, S., and Mao, Y.: Learning from the wisdom of crowds by minimax entropy. In Neural Information Processing Systems (2012), 2204–2212.