Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Presentation:

Kamran Ghasedi

Spring 2015

Outline

- Crowdsourcing
- Problem definition
- K-coin tossing problem
- Markov Decision Process (MDP) and Optimal Policy
- Bayesian Setup
- Accuracy Maximization
- Optimization problem
- Stage-wise expected reward
- Optimistic Knowledge Gradient
- Experiment on simulated data
- Experiment on real dataset
- Conclusion

Paper

Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Dengyong Zhou DENZHO@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

Qiang Liu QLIU1@UCI.EDU
University of California, Irvine, CA 92697

John C. Platt JPLATT@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

Christopher Meek MEEK@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

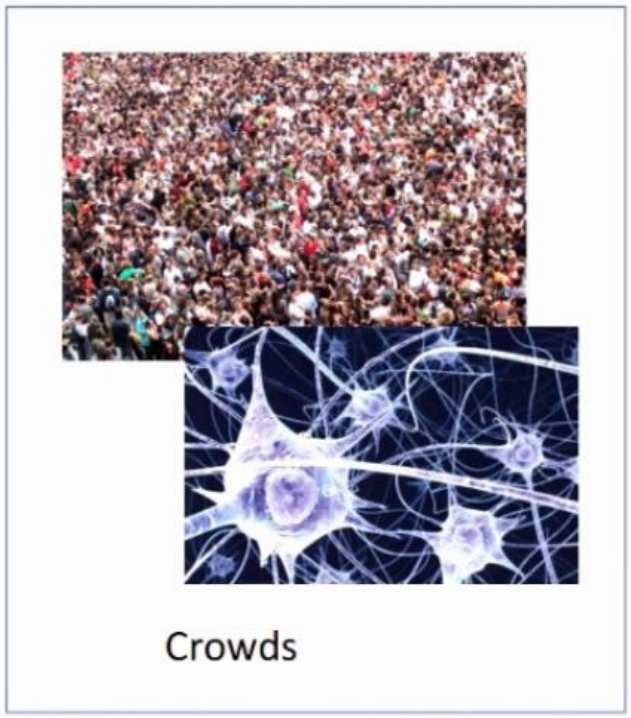
ICML 2014

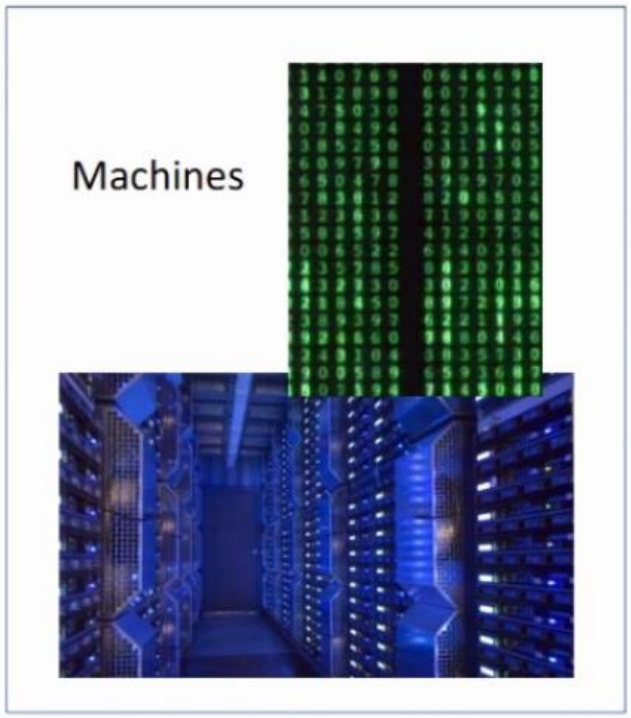
Abstract

We propose a method to aggregate noisy ordinal labels collected from a crowd of workers or annotators. Eliciting ordinal labels is important in tasks such as judging web search quality and rating products. Our method is motivated by the observation that workers usually have difficulty distinguishing between two adjacent ordinal classes whereas distinguishing between two classes which are far away from each other is much easier. We formulate our method as min-

An advanced approach for label aggregation is suggested by Dawid & Skene (1979). They assume that each worker has a latent confusion matrix for labeling. The off-diagonal elements represent the probabilities that a worker mislabels an arbitrary item from one class to another while the diagonal elements correspond to her accuracy in each class. Worker confusion matrices and true labels are jointly estimated by maximizing the likelihood of observed labels. One may further assume a prior distribution over worker confusion matrices and perform Bayesian inference (Raykar et al., 2010; Liu et al., 2012; Chen et al., 2013).

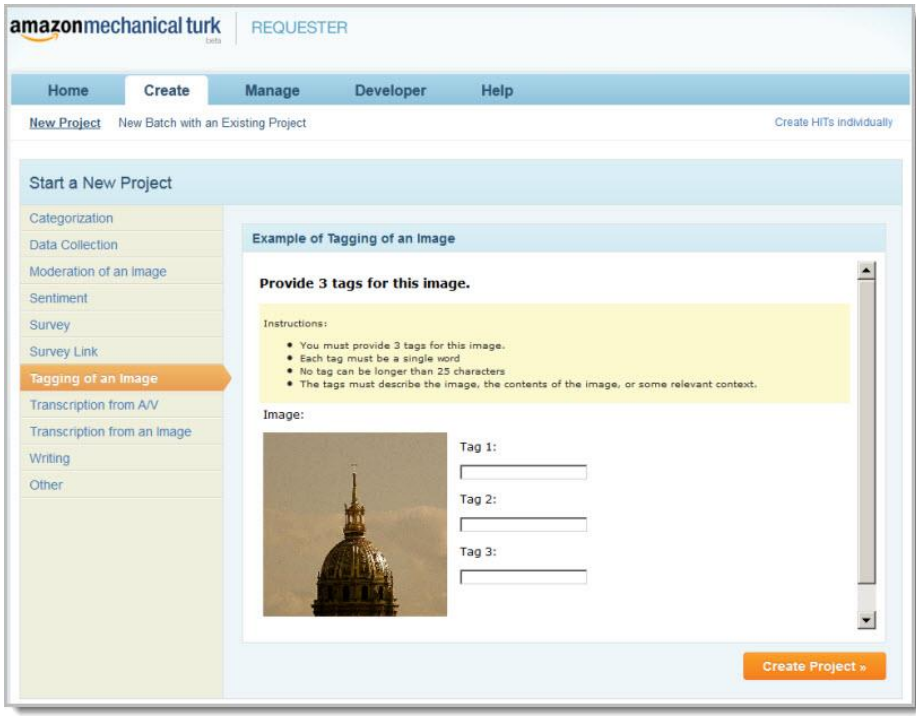
Demand for humans' labels





Crowdsourcing services

Amazon Mechanical Turk



Crowds vs. experts labeling

More data beats clever algorithms!

Noisy labels

Garbage in …

… Garbage out

Crowdsourced labels may be highly noisy..

Toy example

				
	M	O	O	O
	O	O	O	M
	O	M	O	M
	M	M	M	M

Orange (O) vs. Mandarin (M)

Toy example

	M	O		O
	O		O	M
	O	M	O	M
	M	M	M	M

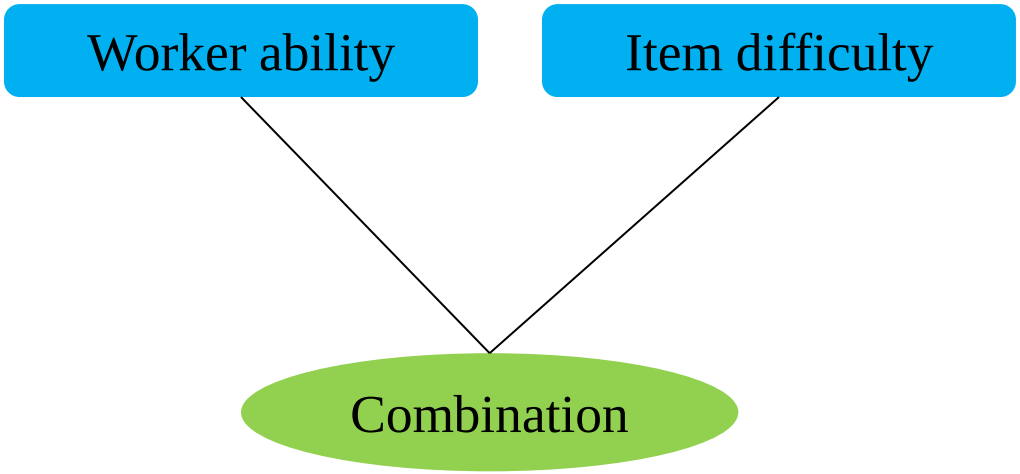
True labels?

Orange (O) vs. Mandarin (M)

Notations

Unobserved true labels: y_j

Assumptions



Roadmap: from multiclass to ordinal

1. Develop a method to aggregate general multiclass labels
2. Adapt general method to ordinal labels

Examples on multiclass labeling

Image categorization

Speech recognition

David & Skene work (1979)

- Assume that each worker has a latent confusion matrix for labeling.
  - The off-diagonal elements represent the probabilities that a worker mislabels an arbitrary item from one class to another.
  - Diagonal elements correspond to its accuracy in each class.

- Worker confusion matrices and true labels are jointly estimated by maximizing the likelihood of observed labels.

- One may further assume a prior distribution over worker confusion matrices and perform Bayesian inference.

Two fundamental tensors

represent an observed confusion from class *c* to class *k* by worker *i* on item *j*

Empirical confusion tensor

$$\widehat{\phi}_{ij}(c,k) = Q(Y_j = c)\mathbb{I}(x_{ij} = k)$$

Expected confusion tensor

$$\phi_{ij}(c,k) = Q(Y_j = c)P(X_{ij} = k|Y_j = c)$$

represent an expected confusion from class *c* to class *k* by worker *i* on item *j*

P: worker label distribution Q: true label distribution

Entropy of the observed labels conditioned on the true labels

Reminder:

$$H(X|Y) = -\sum_{j,c} Q(Y_j = c) \sum_{i,k} P(X_{ij} = k|Y_j = c) \times \log P(X_{ij} = k|Y_j = c).$$

Multiclass maximum conditional entropy

Given the true labels Q, estimate P by

$$\max_{P} \quad H(X|Y)$$

enforce the expected confusion for each worker matches to its empirical confusion.

worker constraints $\sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall i,k,c$

item constraints $\sum \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall j,k,c$

enforce the expected confusion for each item matches to its empirical confusion.

Multiclass minmax conditional entropy

Jointly estimate P and Q by

$$\min_{Q} \max_{P} \quad H(X|Y)$$

subject to

worker constraints $\sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall i,k,c$

item constraints $\sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall j,k,c$

Intuitively: it means that we believe that the observed labels are the least random given the true labels.

Theoretically: minimum conditional entropy can be understood as maximum likelihood.

Lagrangian of maximization problem

$$L = H(X|Y) + L_\sigma + L_\tau + L_\lambda$$

$$L_\sigma = \sum_{i,c,k} \sigma_i(c,k) \sum_j \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right]$$

$$L_\tau = \sum_{j,c,k} \tau_j(c,k) \sum_i \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right]$$

$$L_\lambda = \sum_{i,j,c} \lambda_{ijc} \left[\sum_k P(X_{ij} = k | Y_j = c) - 1 \right]$$

constraints

Probabilistic labeling methd

The (c,k)-th entry represents how likely worker i labels a randomly chosen item in class c as class k.

$$P(X_{ij} = k|Y_j = c) = \frac{1}{Z_{ij}} \exp[\sigma_i(c,k) + \tau_j(c,k)]$$

Z_{ij} normalization factor

worker ability

item difficulty

$$Z_{ij} = \sum_k \exp[\sigma_i(c,k) + \tau_j(c,k)].$$

The (c,k)-th entry represents how likely item j in class c is labeled as class k by a randomly chosen worker.

Dual form of minimax problem

$$\max_{\sigma,\tau,Q} \sum_{j,c} Q(Y_j = c) \sum_i \log P(X_{ij} = x_{ij} | Y_j = c)$$

Lagrangian:

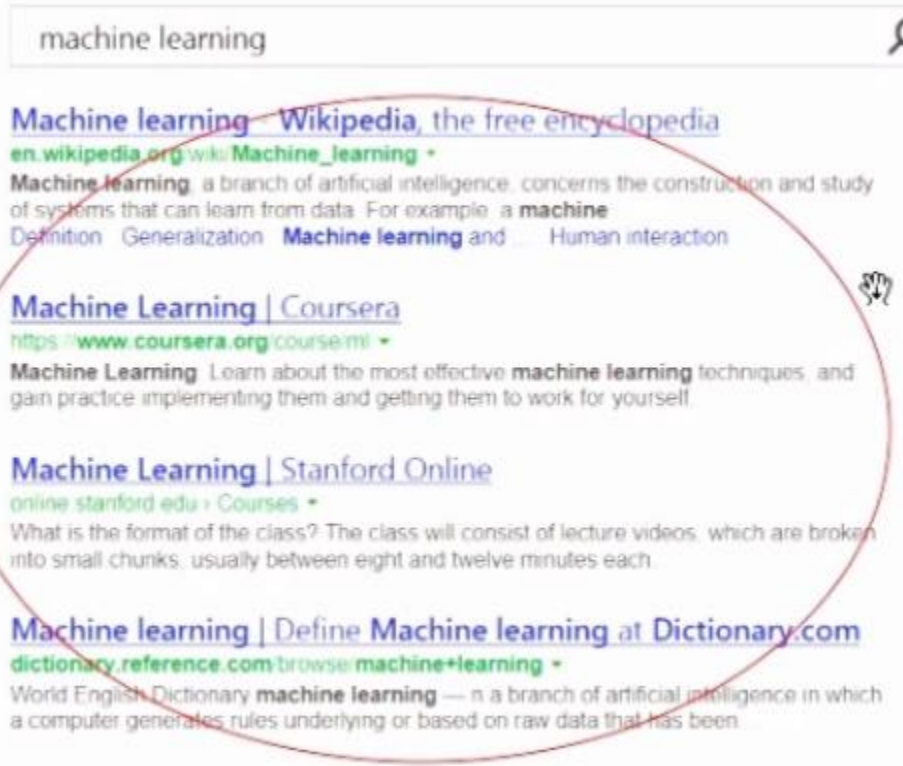
$$\log \left\{ \prod_j \sum_c Q(Y_j = c) \prod_i P(X_{ij} = x_{ij} | Y_j = c) \right\}$$

1. This only generates deterministic labels
2. Equivalent to maximizing complete likelihood

Roadmap: from multiclass to ordinal

1. Develop a method to aggregate general multiclass labels
2. Adapt general method to ordinal labels

Example of ordinal labeling



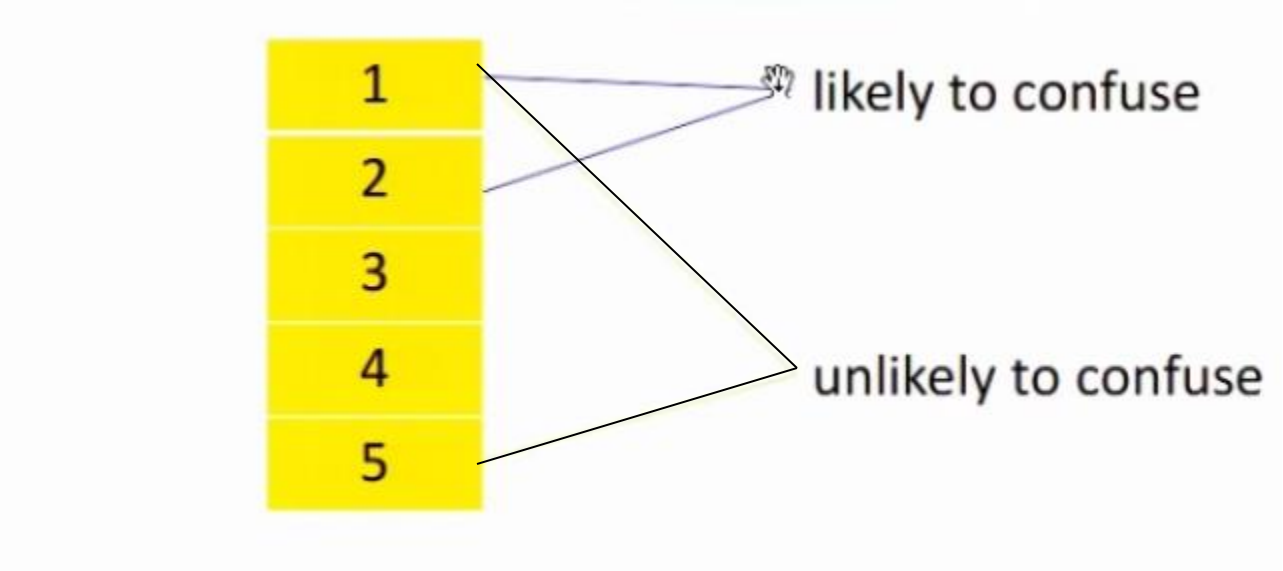
Perfect	1
Excellent	2
Good	3
Fair	4
Bad	5

search results

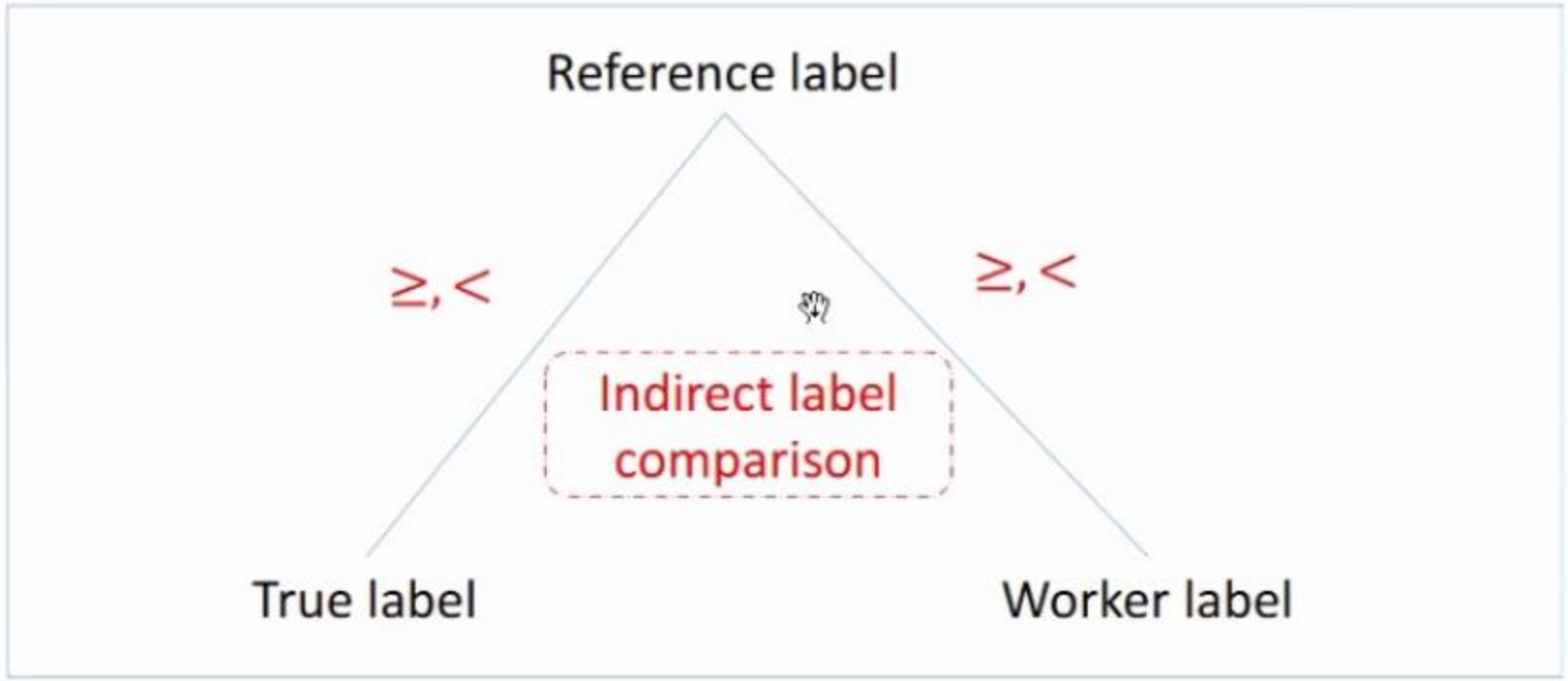
Proceed to ordinal labels

- Formulate assumptions which are specific for ordinal labeling
- Coincide with the previous multiclass method in the case of binary labeling

Assumption for ordinal labeling



Formulating the assumption using pairwise comparison



Ordinal minmax conditional entropy

Jointly estimate P and Q by

$$\min_{Q} \max_{P} \quad H(X|Y)$$

subject to

worker constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right] = 0, \ \forall i, s$$

item constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right] = 0, \ \forall j, s$$

Δ: take on values $<$ or $\geq$

∇: take on values $<$ or $\geq$

Ordinal minmax conditional entropy

Jointly estimate P and Q by

$$\min_{Q} \max_{P} \quad H(X|Y)$$

subject to

worker constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right] = 0, \ \forall i, s$$

item constraints

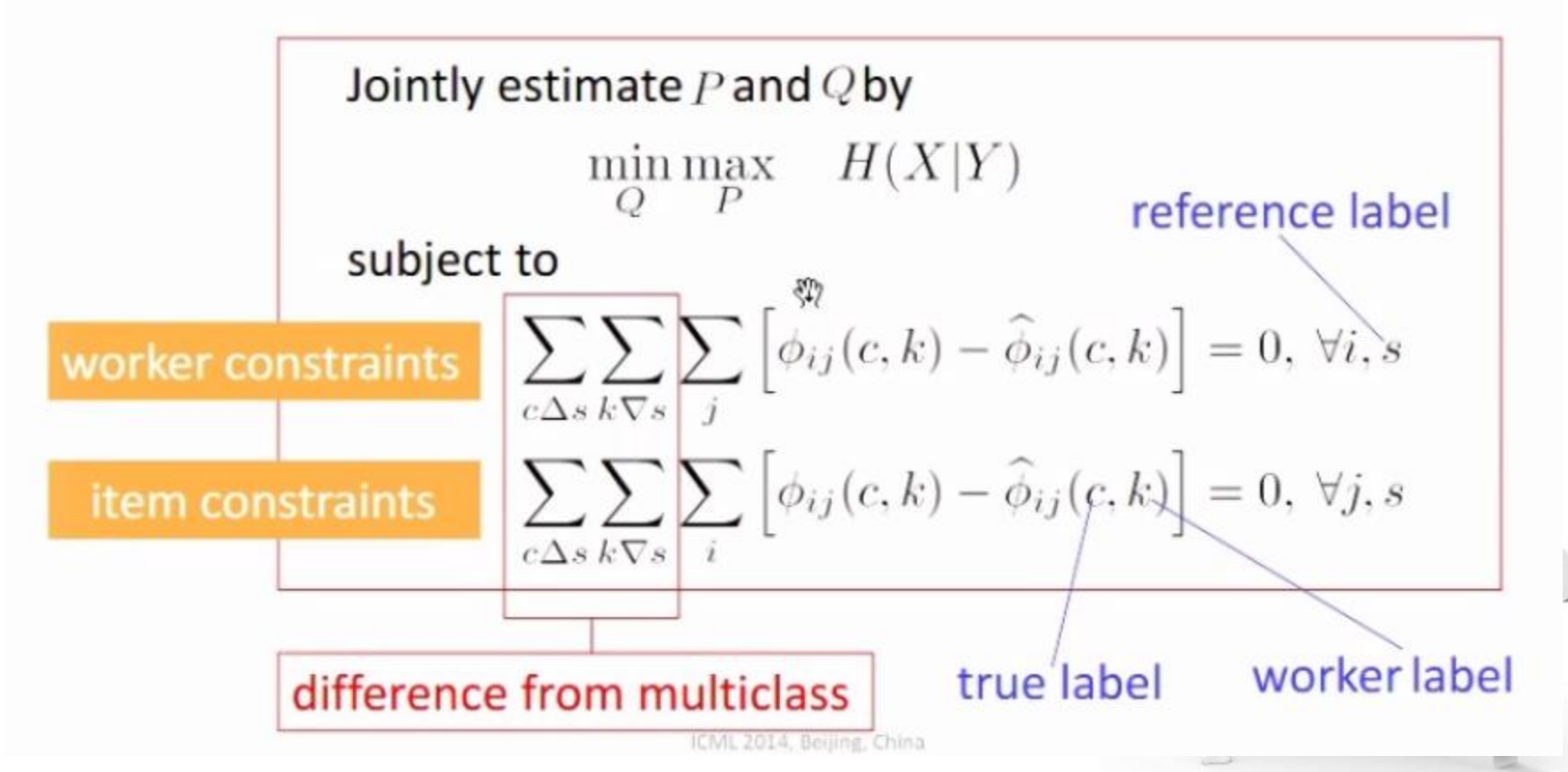
$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right] = 0, \ \forall j, s$$

reference label

true label

worker label

Ordinal minmax conditional entropy



Partition the Cartesian product of the label set

$$\{(c,k)|c<s,k<s\},\ \{(c,k)|c<s,k\geq s\},$$
$$\{(c,k)|c\geq s,k<s\},\ \{(c,k)|c\geq s,k\geq s\}.$$

(a) Partitioning with $s=1$

$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(2,0)$	$(2,1)$	$(2,2)$	$(2,3)$
$(3,0)$	$(3,1)$	$(3,2)$	$(3,3)$

(b) Partitioning with $s=2$

$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(2,0)$	$(2,1)$	$(2,2)$	$(2,3)$
$(3,0)$	$(3,1)$	$(3,2)$	$(3,3)$

(c) Partitioning with $s=3$

$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(2,0)$	$(2,1)$	$(2,2)$	$(2,3)$
$(3,0)$	$(3,1)$	$(3,2)$	$(3,3)$

Explaining the ordinal constraints

For example, let $\Delta = <, \nabla = \geq$:

$$\sum_{c<s} \sum_{k \geq s} \hat{\phi}_{ij}(c,k) = Q(Y_j < s)\mathbb{I}(x_{ij} \geq s)$$

counting mistakes in ordinal sense

Probabilistic rating model

By the KKT conditions, the dual problem leads to

$$P(X_{ij} = k | Y_j = c) = \frac{1}{Z_{ij}} \exp[\sigma_i(c,k) + \tau_j(c,k)]$$

worker ability

$$\sigma_i(c,k) = \sum_{s \geq 1} \sum_{\Delta, \nabla} \sigma_{is}^{\Delta, \nabla} \mathbb{I}(c\Delta s, k \nabla s)$$

item difficulty

$$\tau_j(c,k) = \sum_{s \geq 1} \sum_{\Delta, \nabla} \tau_{js}^{\Delta, \nabla} \mathbb{I}(c\Delta s, k \nabla s)$$

structured

Regularization

Two goals:

1. Prevent over fitting
2. Fix the deterministic label issue to generate probabilistic labels

Regularized minmax conditional entropy

Jointly estimate P and Q by

$$\min_{Q} \max_{P} \; H(X|Y) \quad + \text{regularization terms}$$

subject to

worker constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right] \approx 0, \; \forall i, s$$

item constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k) \right] \approx 0, \; \forall j, s$$

Regularized minmax conditional entropy

Jointly estimate P and Q by

$$\min_Q \max_P \; H(X|Y) - H(Y) - \frac{1}{\alpha}\Omega(\xi) - \frac{1}{\beta}\Psi(\zeta)$$

subject to

worker constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_j \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = \xi_{is}^{\Delta,\nabla}$$

item constraints

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_i \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = \zeta_{js}^{\Delta,\nabla}$$

$$H(Y) = -\sum_{j,c} Q(Y_j = c) \log Q(Y_j = c).$$

Dual problem

$$\max_{\sigma,\tau,Q} \sum_{j,c} Q(Y_j = c) \sum_i \log P(X_{ij} = x_{ij} | Y_j = c) + H(Y) - \alpha\Omega(\sigma) - \beta\Psi(\tau)$$

1. This generates probabilistic labels
2. Equivalent to maximizing marginal likelihood

Choosing regularization parameters

- Cross-validation: 5 or 10 folds
- Random split
- Compare the likelihood of worker labels

Don’t need ground truth labels for cross-validation!

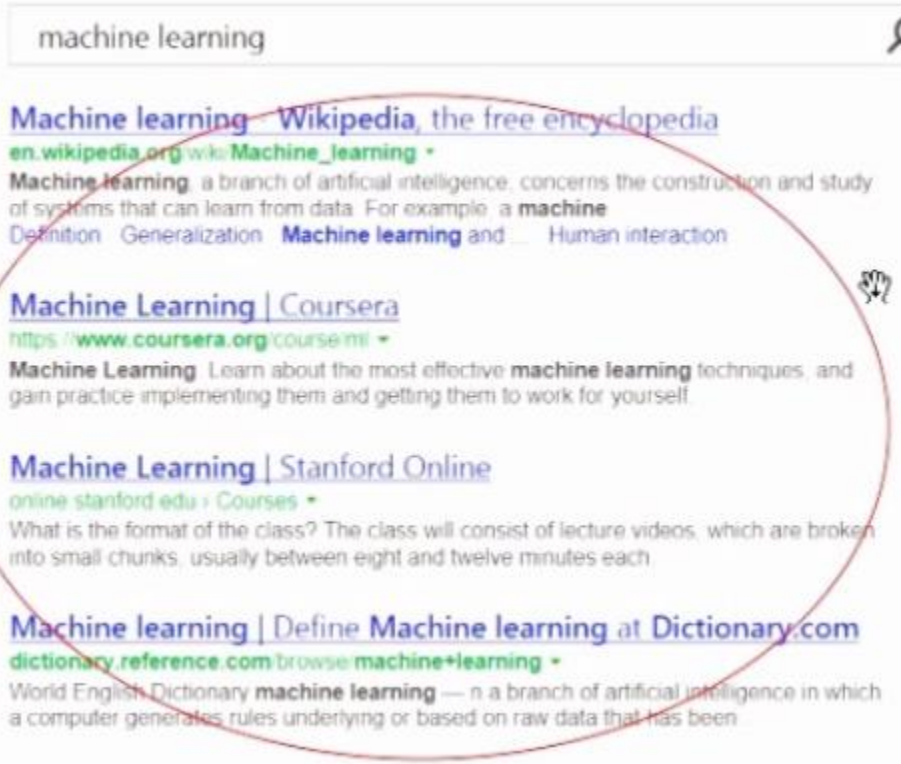
Experiments: metric

- Evaluation metrics
  - L0 error: $\mathrm{L0} = \mathbb{I}(y \neq \widehat{y})$
  - L1 error: $\mathrm{L1} = |y - \widehat{y}|$
  - L2 error: $\mathrm{L2} = |y - \widehat{y}|^2$

Experiments: baselines

- Compare regularized minimax condition entropy to
  - Majority voting
  - Dawid-Skene method (1979, see also its Bayesian version in Raykar et al. 2010, Liu et al. 2012, Chen at al. 2013)
  - Latent trait analysis (Andrich 1978, Master 1982, Uebersax and Grove 1993, Mineiro 2011)

Web search data



Perfect	1
Excellent	2
Good	3
Fair	4
Bad	5

search results

Web search data

- Some facts about the data:
  - 2665 query-URL pairs and a relevance rating scale from 1 to 5
  - 177 non-expert workers with average error rate 63%
  - Each query-URL pair is judged by 6 workers
  - True labels are created via consensus from 9 experts
  - Dataset created by Gabriella Kazai of Microsoft

Result for web search data

	L0 Error	L1 Error	L2 Error
Majority vote	0.269	0.428	0.930
Dawid & Skene	0.170	0.205	0.539
Latent trait	0.201	0.211	0.481
Entropy multiclass	0.111	0.131	0.419
Entropy ordinal	0.104	0.118	0.384

Probabilistic labels vs. error rate

Price prediction data

$0 – $50	1
$51 – $100	2
$101 – $250	3
$251 – $500	4
$501 – $1000	5
$1001 – $2000	6
$2001 – $5000	7

Price prediction data

- Some facts about the data:
  - 80 household items collected from stores like Amazon and Costco
  - Prices predicted by 155 students of UC Irvine
  - Average error rate 69% and systematically biased
  - Dataset created by Mark Steyvers of UC Irvine

Result for price prediction data

	L0 Error	L1 Error	L2 Error
Majority vote	0.675	1.125	1.605
Dawid & Skene	0.650	1.050	1.517
Latent trait	0.688	1.063	1.504
Entropy multiclass	0.675	1.150	1.643
Entropy ordinal	0.613	0.975	1.492

Conclusion

- Minmax conditional entropy principle for crowdsourcing
- Adjacency confusability assumption in ordinal labeling
- Ordinal labeling model with structured confusion matrices

Thanks for your attention