Algorithmic Crowdsourcing

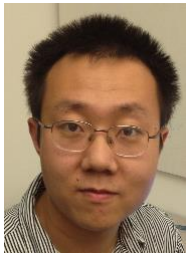
Denny Zhou

Microsoft Research Redmond

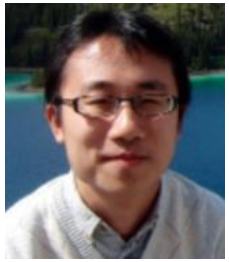
Dec 9, NIPS13, Lake Tahoe

Main collaborators

John Platt
(Microsoft)

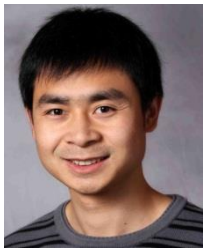


Xi Chen
(UC Berkeley)



Chao Gao
(Yale)

Nihar Shah
(UC Berkeley)

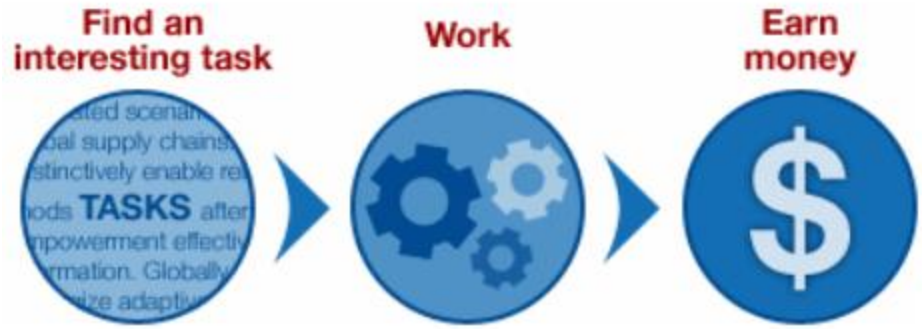


Qiang Liu
(UC Irvine)

Machine learning + crowdsourcing

- Almost all machine learning applications need training labels
- By crowdsourcing we can obtain many labels in a short time at very low cost

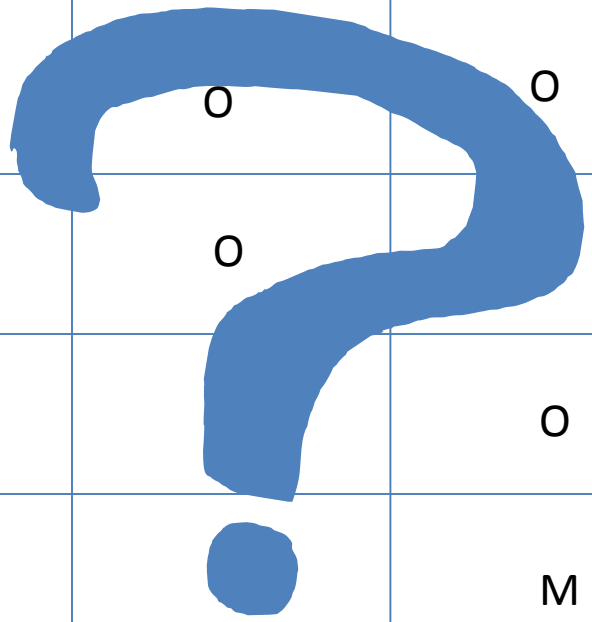
amazonmechanical turk
beta
Artificial Artificial Intelligence



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

Repeated labeling: Orange (O) vs. Mandarin (M)

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



Repeated labeling: Orange (O) vs. Mandarin (M)

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

Need assumptions!

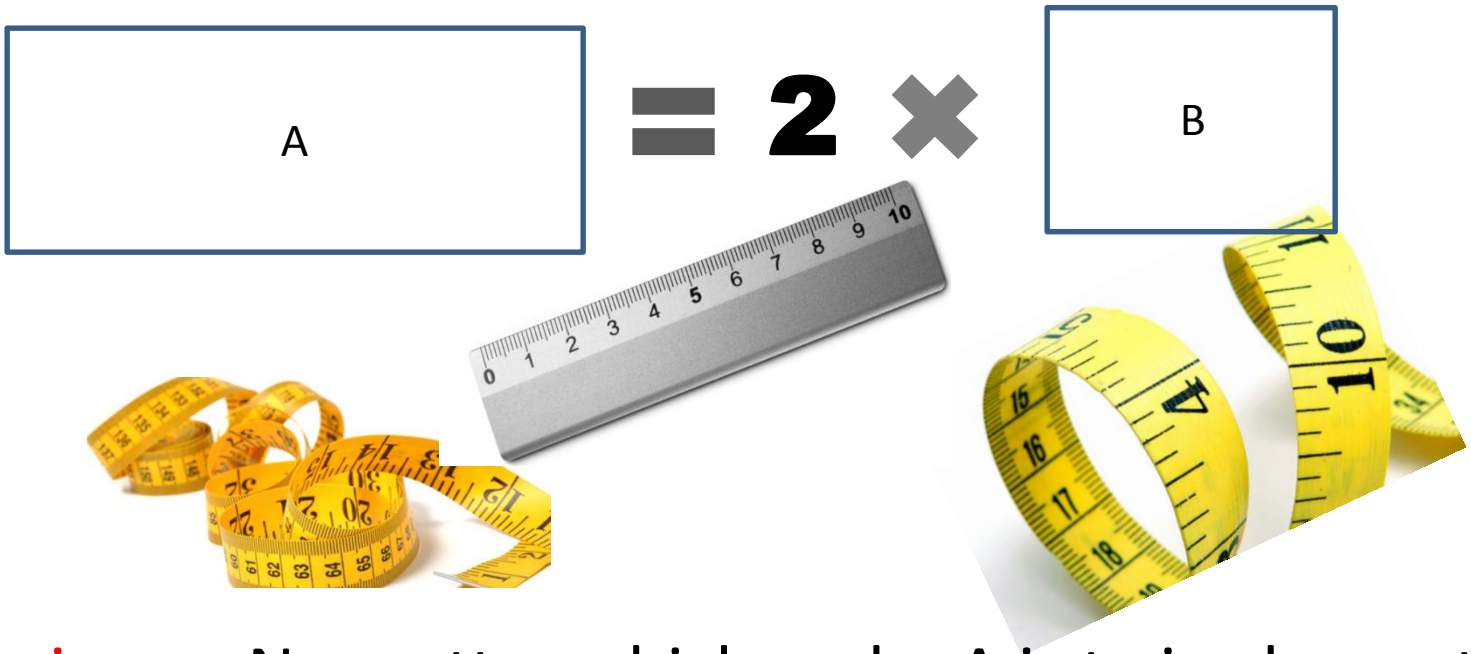
Repeated labeling: Orange (O) vs. Mandarin (M)

How to make assumptions?

- Intuitively, label quality depends on worker ability and item difficulty. But,
  - How to measure worker ability?
  - How to measure item difficulty?
  - How to combine worker ability and item difficulty?
  - How to infer worker ability and item difficulty?
  - How to infer ground truth?

A single assumption for all questions

Our assumption: measurement objectivity



Invariance: No matter which scale, A is twice larger than B

A single assumption for all questions

Our assumption: measurement objectivity

A  = 2 ×  B

How to formulate invariance for mental measuring?

A single assumption for all questions

Our assumption: measurement objectivity

A  = 2 ×  B

Assume a set $\aleph$ of equally difficult questions:

R_A: number of right answers
W_A: number of wrong answers

$$\frac{R_A/W_A}{R_B/W_B}$$

A single assumption for all questions

Our assumption: measurement objectivity

A  = 2 ×  B

Assume another set $\aleph'$ of equally difficult questions:

R'_A: number of right answers
W'_A: number of wrong answers

$$\frac{R'_A / W'_A}{R'_B / W'_B}$$

A single assumption for all questions

Our assumption: measurement objectivity

A  = 2 ×  B

$$\frac{R_A/W_A}{R_B/W_B} = \frac{R'_A/W'_A}{R'_B/W'_B}$$

A single assumption for all questions

Our assumption: measurement objectivity

A  = 2 ×  B

For multiclass labeling, we count the number of misclassifications from one class to another

Measurement objectivity assumption leads to a unique model!

worker i, item j
labels c, k

data matrix X_{ij}
true label Y_j

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

worker confusion matrix

item confusion matrix

Estimation procedure

- First, estimate the worker and item confusion matrices by maximizing marginal likelihood
- Then, estimate the labels by using Bayes' rule with the estimated confusion matrices

Two steps can be seamlessly unified in EM!

Expectation-Maximization (EM)

- Initialize label estimates via majority vote
- Iterate till converge:
  - Given the estimates of labels, estimate worker and item confusion matrices
  - Given the estimates of worker and item confusion matrices, estimate labels

Prevent overfitting

- Equivalently formulate our solution into minimax conditional entropy
- Prevent overfitting by a natural regularization

Minimax conditional entropy

- True label distribution: $Q(Y)$
- Define two 4-dim tensors
  - Empirical confusion tensor

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

  - Expected confusion tensor

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

Minimax conditional entropy

- Jointly estimate P and Q by

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

subject to

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

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

Minimax conditional entropy

- Exactly recover the previous model via the dual of maximum entropy

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

Nothing but Lagrangian multipliers!

- Estimate true labels by minimizing maximum entropy (= maximizing likelihood)

Regularization

- Move from exact matching to approximate matching:

$$\sum_j \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] \approx 0, \forall i,k,c$$

$$\sum_i \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] \approx 0, \forall j,k,c$$

Regularization

- Move from exact matching to approximate matching:

$$\sum_j \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = \xi_i(c,k), \ \forall i,k,c,$$

$$\sum_i \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = \zeta_j(c,k), \ \forall j,k,c$$

- Penalize large fluctuations:

$$\min_Q \max_P \ H(X|Y) - \alpha\|\xi\|^2 - \beta\|\zeta\|^2$$

Experimental results

- Bluebirds data (error rates)

# Workers	10	15	20	25
Minimax Entropy	0.150 ±0.061	0.122 ±0.026	0.097 ±0.017	0.090 ±0.016
Dawid & Skene	0.142 ±0.039	0.132 ±0.025	0.114 ±0.017	0.117 ±0.017
Belief Propagation	0.143 ±0.040	0.133 ±0.026	0.117 ±0.018	0.121 ±0.019
Majority Vote	0.244 ±0.065	0.234 ±0.052	0.240 ±0.034	0.242 ±0.030

  - Belief propagation: Variational Inference for Crowdsourcing (Liu et al. NIPS 2013)
  - Other methods in the literature cannot outperform Dawid & Skene (1979)
  - Some are even worse than majority voting
  - Data: The multidimensional wisdom of crowds (Welinder et al, NIPS 2010)

Experimental results

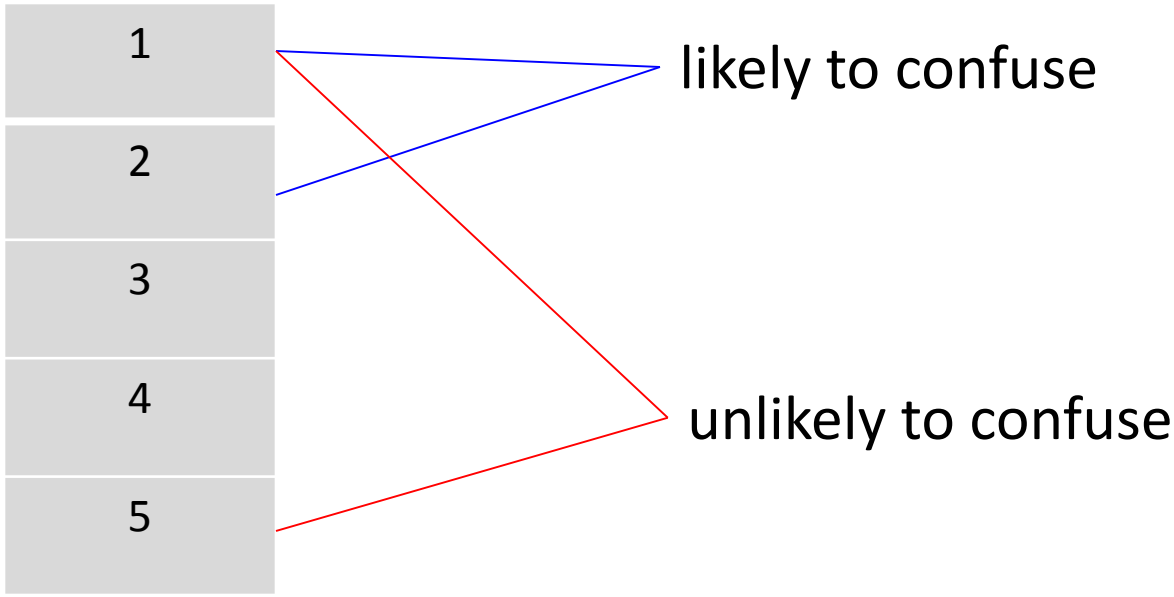
- Web search data (error rates)

	Majority Vote	Dawid & Skene	Latent Trait	Minimax Entropy
L0 Error	0.269	0.170	0.201	0.111
L1 Error	0.428	0.205	0.211	0.131
L2 Error	0.930	0.539	0.481	0.419

- Latent trait analysis code from: http://www.machinedlearnings.com
- Data: Learning from the wisdom of crowds by minimax entropy (Zhou et al., NIPS 2012)

Crowdsourced ordinal labeling

- Ordinal labels: web search, product rating
- Our assumption: adjacency confusability



Ordinal minimax conditional entropy

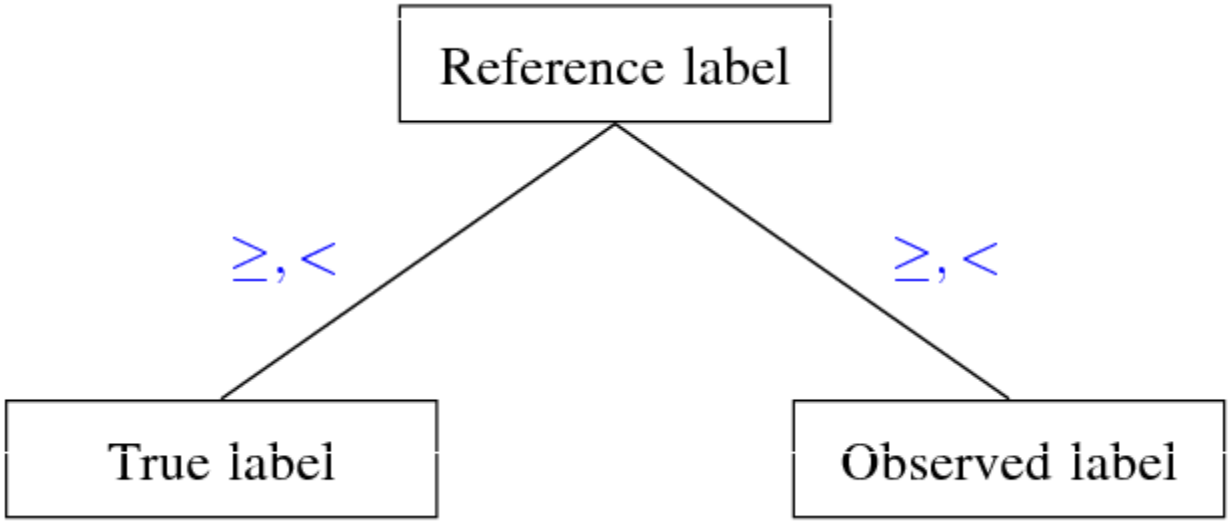
- Minimax conditional entropy with the ordinal-based worker and item constraints:

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

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

for all Δ and ∇ taking values from $\{\geq, <\}$

Constraints: indirect label comparison



Ordinal labeling model

- Obtain the same model except confusion matrices are subtly structured by

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

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

- Fewer parameter, less model complexity

Experimental results

Web search data

	Majority Vote	Dawid & Skene	Latent Trait	Entropy (M)	Entropy (O)
L0 Error	0.269	0.170	0.201	0.111	**0.104**
L1 Error	0.428	0.205	0.211	0.131	**0.118**
L2 Error	0.930	0.539	0.481	0.419	**0.384**

- Latent trait analysis code from: http://www.machinedlearnings.com
- Entropy(M): regularized minimax conditional entropy for multiclass labels
- Entropy(O): regularized minimax conditional entropy for ordinal labels
- Data: Learning from the wisdom of crowds by minimax entropy (Zhou et al., NIPS 2012)

Experimental results

Price estimation data

	Majority Vote	Dawid & Skene	Latent Trait	Entropy (M)	Entropy (O)
L0 Error	0.675	0.650	0.688	0.675	**0.613**
L1 Error	1.125	1.050	1.063	1.150	**0.975**
L2 Error	1.605	1.517	1.504	1.643	**1.492**

- Latent trait analysis code from: http://www.machinedlearnings.com
- Entropy(M): regularized minimax conditional entropy for multiclass labels
- Entropy(O): regularized minimax conditional entropy for ordinal labels
- Data: 7 price ranges from least expensive to most expensive (Liu et al. NIPS 13)

Why latent trait model doesn't work?

When solving a given problem try to avoid solving a more general problem as an intermediate step.

-Vladimir Vapnik

ordinal label = score range

Error bounds: problem setup

- Observed Data: $X = (X_{ij})_{I \times J}$
- Unknown true labels: $Y = (Y_1, ..., Y_J)$
- Unknown workers' accuracies: $p = (p_1, ..., p_I)$
- Simplified model of Dawid and Skene (1979) and minimax conditional entropy

Dawid-Skene estimator (1979)

- Complete likelihood: $\mathbb{P}(X, Y|p)$
- Marginal likelihood: $\mathbb{P}(X|p)$
- Estimating workers' accuracy by

$$\hat{p} = \arg\max_{p} \mathbb{P}(X|p)$$

- Estimating true labels by (plug-in)

$$\hat{z}_j = \mathbb{P}(Y_j = 1|X, \hat{p})$$

Theorem (lower bound)

For any estimator, there exists a least favorable $p \in \mathcal{P}_{q,\bar{\mu}}$ (parameter space)

$$\frac{1}{J}\sum_j |\hat{z}_j - Y_j| \gtrsim \exp\left(-8I \max\left\{2q, \frac{1}{2}D(\bar{\mu}||1-\bar{\mu})\right\}\right)$$

$$\mathcal{P}_{q,\bar{\mu}} = \left\{ p \in \mathbb{R}^I : \frac{1}{I}\sum_i (2p_i - 1)^2 = q, \frac{1}{I}\sum_i p_i = \bar{\mu} \right\}$$

Theorem (upper bound)

Under mild assumptions, Dawid-Skene estimator is optimal in Wald's sense

$$\frac{1}{J}\sum_{j}|\hat{z}_j - Y_j| \lesssim \exp\left(-\frac{1}{4}I\max\left\{2q, \frac{1}{2}D(\bar{\mu}||1-\bar{\mu})\right\}\right)$$

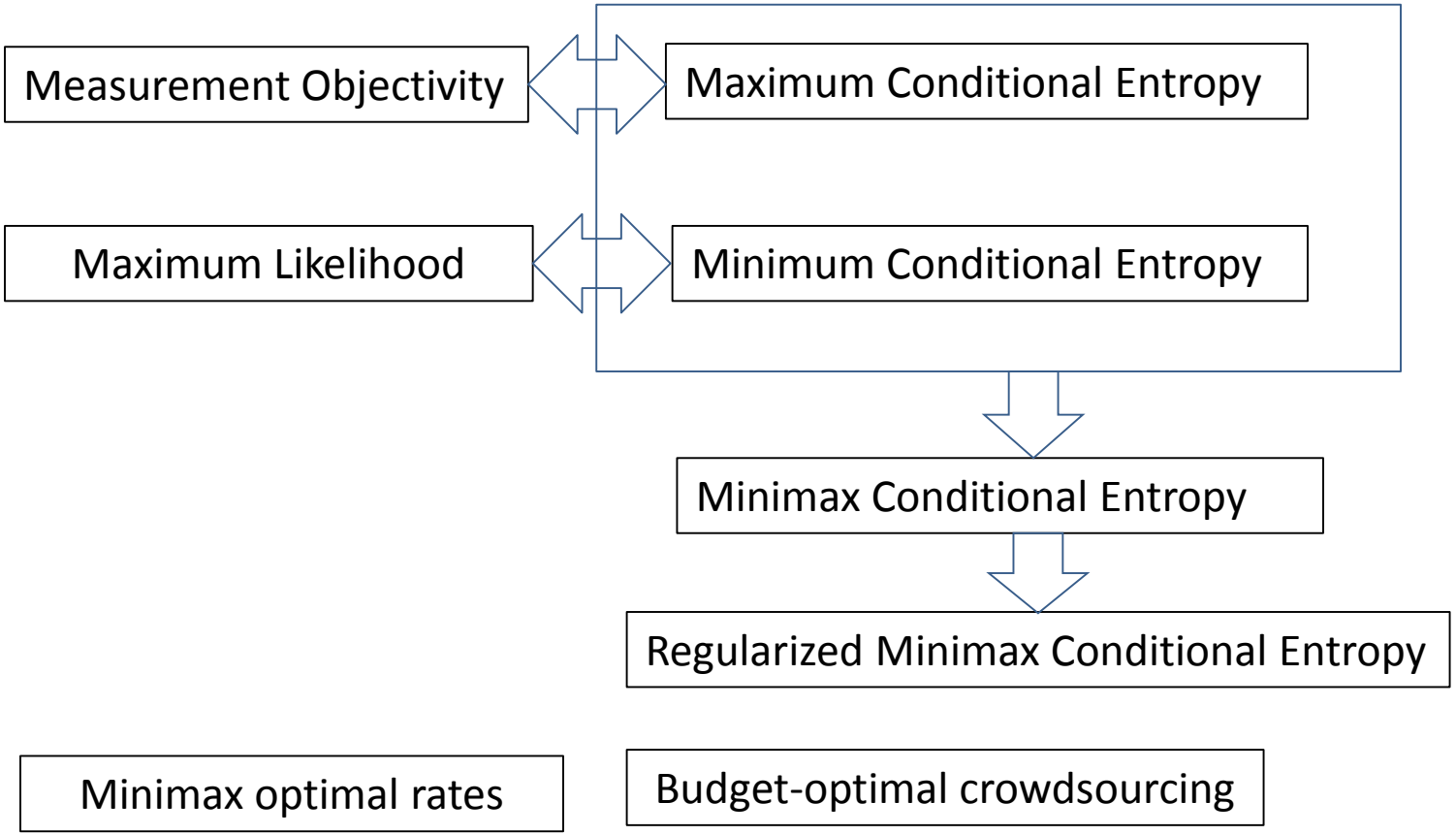
Budget-optimal crowdsourcing

We propose the following formulation:

- Given n biased coins, we want to know which are biased to heads and which are biased to tails
- We have a budget of tossing m times in total
- Our goal is to maximize the accuracy of prediction based on observed tossing outcomes

Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourcing (Chen et al, ICML 2013)

Summary


Measurement Objectivity
Maximum Conditional Entropy
Maximum Likelihood
Minimum Conditional Entropy
Minimax Conditional Entropy
Regularized Minimax Conditional Entropy
Minimax optimal rates
Budget-optimal crowdsourcing

- Chao Gao and Dengyong Zhou. Minimax Optimal Convergence Rates for Estimating Ground Truth from Crowdsourced Labels, no. MSR-TR-2013-110, October 2013
- Dengyong Zhou, Qiang Liu, John Platt and Chris Meek. Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy. Submitted.
- Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourcing, in *Proceedings of the 30th International Conference on Machine Learning* (ICML), 2013
- Dengyong Zhou, John Platt, Sumit Basu, and Yi Mao. Learning from the Wisdom of Crowds by Minimax Entropy, in *Advances in Neural Information Processing Systems* (NIPS), December 2012

http://research.microsoft.com/en-us/projects/crowd/