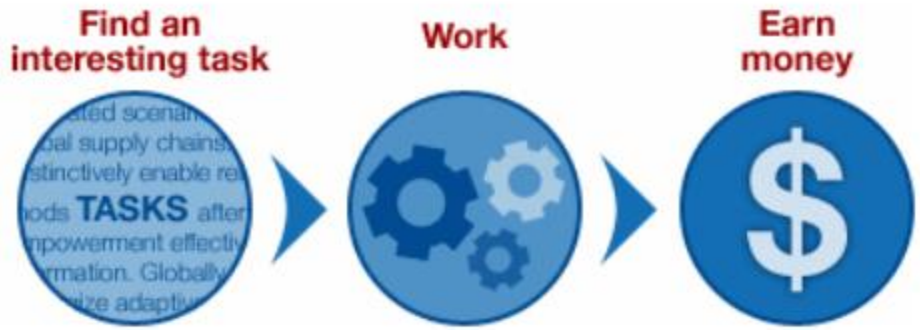
Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Denny Zhou Qiang Liu John Platt Chris Meek

ICML | Beijing

Microsoft

amazon mechanical turk
Artificial Artificial Intelligence



Crowds vs experts labeling: strength

Time saving

Money saving

Big labeled data

More data beats cleverer algorithms

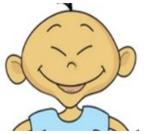
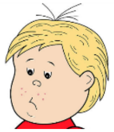
Crowds vs experts labeling: weakness

Garbage in …

… Garbage out

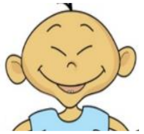
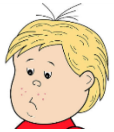
Crowdsourced labels may be highly noisy

Non-experts, redundant labels

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

Orange (O) vs. Mandarin (M)

Non-experts, redundant labels

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

True labels?

Orange (O) vs. Mandarin (M)

Items

Workers

Workers	1	2	…	j	…
1	x_{11}	x_{12}	…	x_{1j}	…
2	x_{21}	x_{21}	…	x_{2j}	…
…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…
…	…	…	…	…	…

Observed worker labels

Unobserved true labels: y_j

Roadmap: from multiclass to ordinal

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

Examples on multiclass labeling

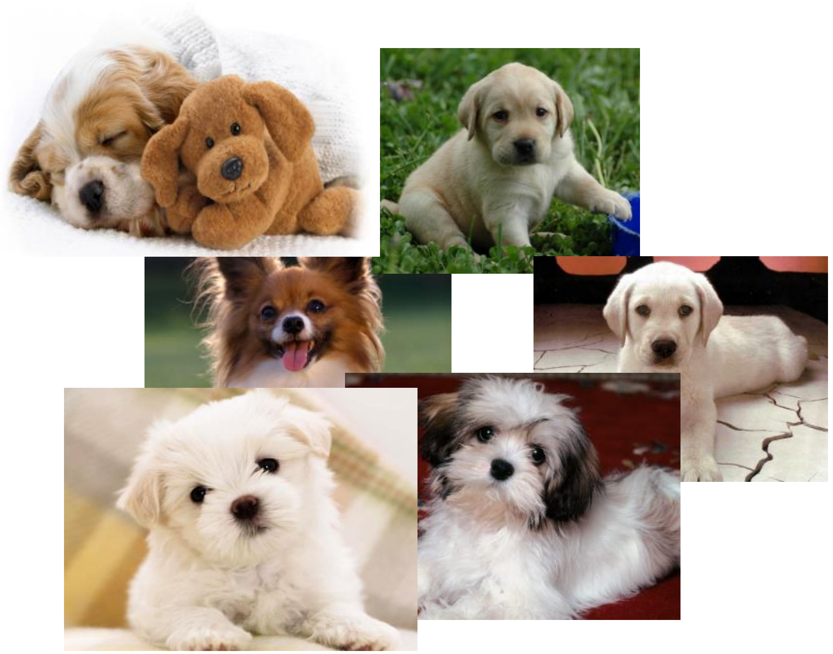


Image categorization

Speech recognition

Introduce two fundamental concepts

Empirical count of wrong/correct labels

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

Expected number of wrong/correct labels

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

P: worker label distribution Q: true label distribution

Multiclass maximum conditional entropy

Given the true labels Q, estimate P by

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

subject to

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

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

Multiclass minimax conditional entropy

Jointly estimate P and Q by

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

subject to

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

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

Lagrangian dual

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

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

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

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

constraints

Probabilistic labeling model

By the optimization theory, the dual problem leads to

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

Z_{ij} normalization factor

worker ability

item difficulty

Dual problem

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

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 the general method to ordinal labels

An example on ordinal labeling

machine learning

[Machine learning - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Machine_learning
Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. For example, a **machine** ...
Definition · Generalization · **Machine learning** and ... · Human interaction

[Machine Learning | Coursera](#)
https://www.coursera.org/course/ml
Machine Learning. Learn about the most effective **machine learning** techniques, and gain practice implementing them and getting them to work for yourself.

[Machine Learning | Stanford Online](#)
online.stanford.edu › Courses
What is the format of the class? The class will consist of lecture videos, which are broken into small chunks, usually between eight and twelve minutes each.

[Machine learning | Define Machine learning at Dictionary.com](#)
dictionary.reference.com/browse/machine+learning
World English Dictionary **machine learning** — n a branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been ...

search results

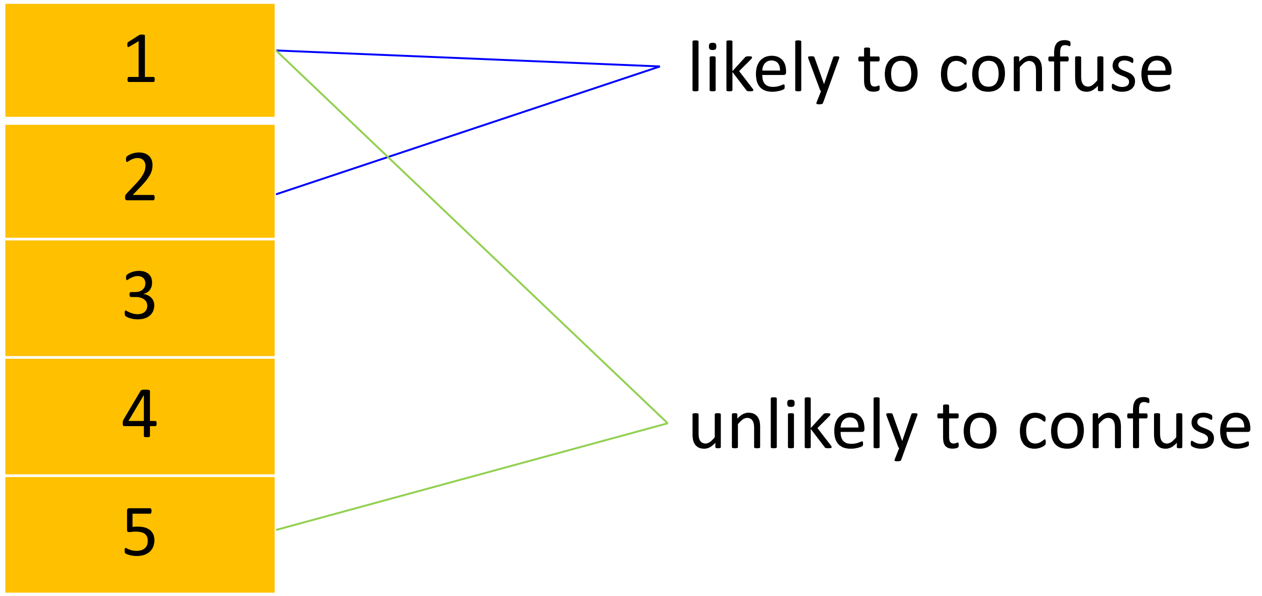
Perfect	1
Excellent	2
Good	3
Fair	4
Bad	5

To proceed to ordinal labels

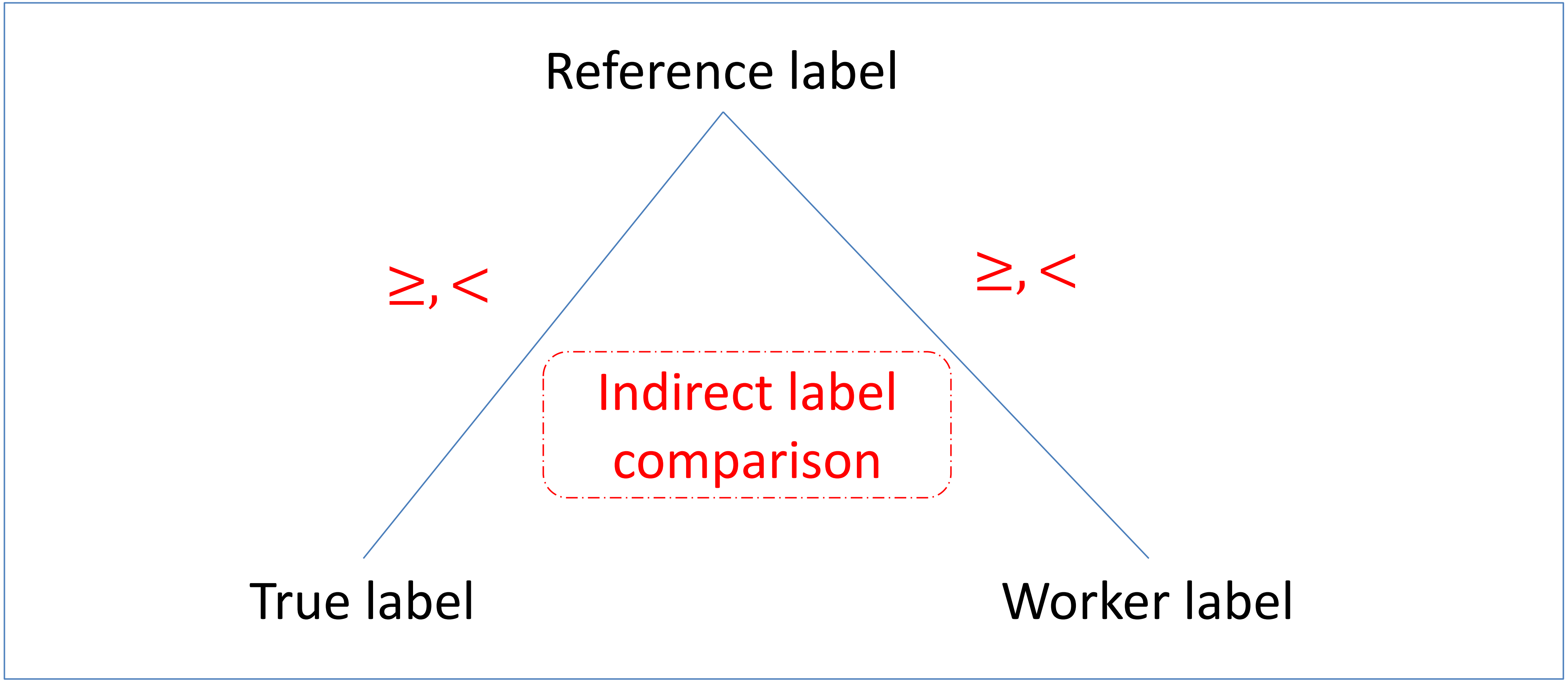
- Formulate assumptions which are specific for ordinal labeling
- Coincide with the previous multiclass method in the case of binary labeling

Our assumption for ordinal labeling

adjacency confusability



Formulating this assumption though pairwise comparison



Ordinal minimax 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 minimax conditional entropy

Jointly estimate P and Q by

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

reference label

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

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

true label worker label

Ordinal minimax 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

difference from multiclass

true label

worker label

Explaining the ordinal constraints

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

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

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 minimax conditional entropy

Jointly estimate P and Q by

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

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 minimax conditional entropy

Jointly estimate P and Q by

$$\min_{Q} \max_{P} \; H(X|Y) - \boxed{H(Y) - \frac{1}{\alpha}\Omega(\xi) - \frac{1}{\beta}\Psi(\zeta)}$$

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

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

Dual problem

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

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!

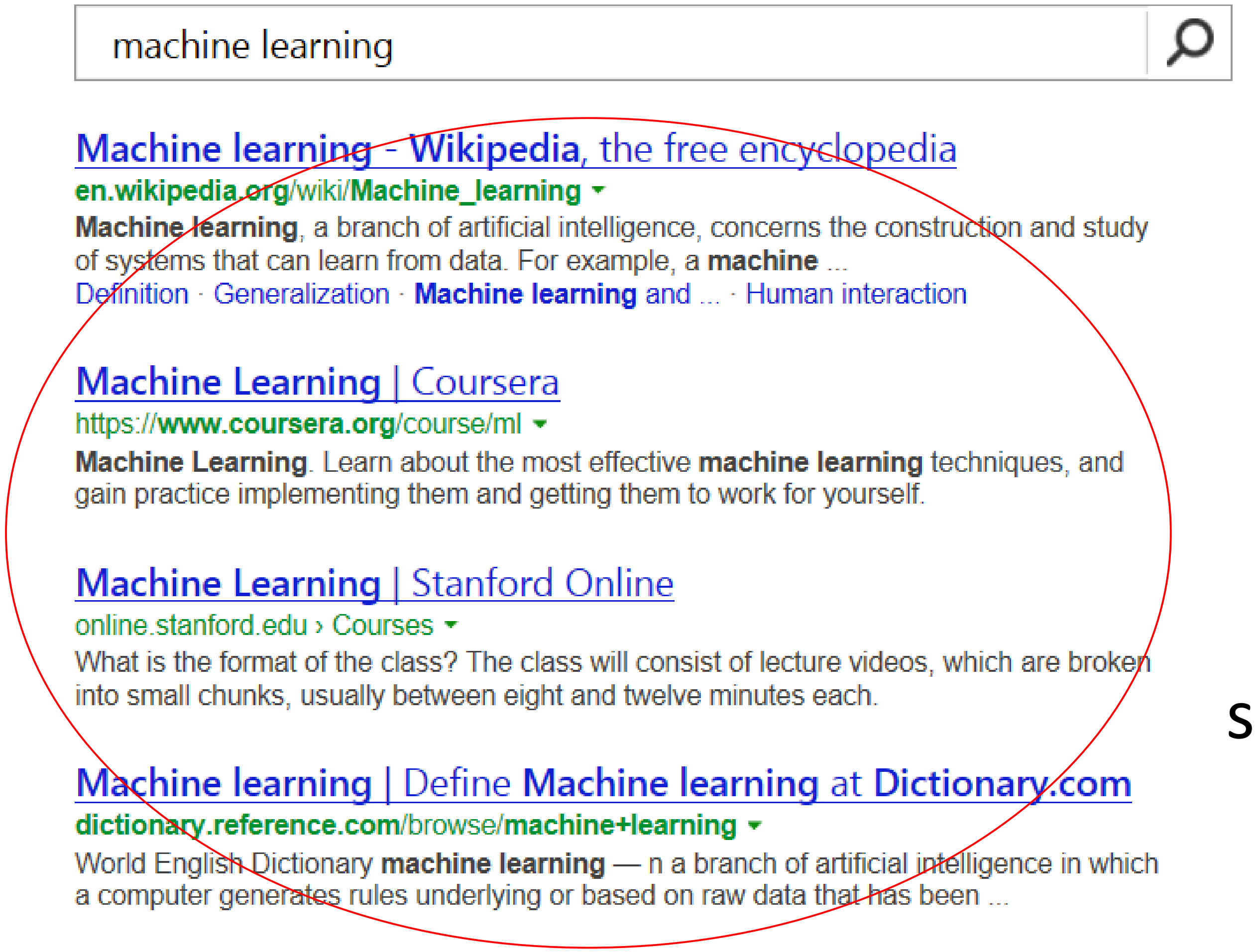
Experiments: metrics

- 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


machine learning

Machine learning - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Machine_learning
Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. For example, a **machine** ...
Definition · Generalization · **Machine learning** and ... · Human interaction

Machine Learning | Coursera
https://www.coursera.org/course/ml
Machine Learning. Learn about the most effective **machine learning** techniques, and gain practice implementing them and getting them to work for yourself.

Machine Learning | Stanford Online
online.stanford.edu › Courses
What is the format of the class? The class will consist of lecture videos, which are broken into small chunks, usually between eight and twelve minutes each.

Machine learning | Define Machine learning at Dictionary.com
dictionary.reference.com/browse/machine+learning
World English Dictionary **machine learning** — n a branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been ...

search results

Perfect	1
Excellent	2
Good	3
Fair	4
Bad	5

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

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 rates

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

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

Summary

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

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