Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Denny Zhou      Qiang Liu      John Platt      Chris Meek
Crowds vs experts labeling: strength

Time saving

Money saving

Big labeled data

More data beats cleverer algorithms
Crowds vs experts labeling: weakness

Crowdsourced labels may be highly noisy
Non-experts, redundant labels

<table>
<thead>
<tr>
<th></th>
<th>Orange (O) vs. Mandarin (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-experts</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>
Non-experts, redundant labels

<table>
<thead>
<tr>
<th></th>
<th>Orange (O)</th>
<th>Mandarin (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>O</td>
</tr>
<tr>
<td>2</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>3</td>
<td>O</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

Orange (O) vs. Mandarin (M)
### Workers

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$j$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>...</td>
<td>$x_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>$x_{21}$</td>
<td>$x_{21}$</td>
<td>...</td>
<td>$x_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_{i1}$</td>
<td>$x_{i2}$</td>
<td>...</td>
<td>$x_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**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

\[ \hat{\phi}_{ij}(c, k) = Q(Y_j = c) \mathbb{I}(x_{ij} = k) \]

**Expected** number of wrong/correct labels

\[ \phi_{ij}(c, k) = Q(Y_j = c) P(X_{ij} = k \mid Y_j = c) \]

*P*: worker label distribution  \quad *Q*: true label distribution
Multiclass maximum conditional entropy

Given the true labels $Q$, estimate $P$ by

$$\max_P H(X|Y)$$

subject to

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

- **item constraints**
  $$\sum_i \left[ \phi_{ij}(c,k) - \hat{\phi}_{ij}(c,k) \right] = 0, \forall j, k, c$$
Jointly estimate $P$ and $Q$ by

$$\min_Q \max_P H(X|Y)$$

subject to

**worker constraints**

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

**item constraints**

$$\sum_i \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \ \forall j, k, c$$
Lagrangian dual

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

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

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

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

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

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

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

<table>
<thead>
<tr>
<th>Search Result</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning - Wikipedia, the free encyclopedia</td>
<td>Perfect 1</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Excellent 2</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Good 3</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Fair 4</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Bad 5</td>
</tr>
</tbody>
</table>

search results
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

likely to confuse

unlikely to confuse
Formulating this assumption through pairwise comparison

Reference label

≥, <

True label

≥, <

Worker label

Indirect label comparison
Ordinal minimax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_Q \max_P H(X|Y)$$

subject to

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

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

$\Delta$: take on values $< \text{ or } \geq$

$\nabla$: take on values $< \text{ or } \geq$
Jointly estimate $P$ and $Q$ by

$$
\min \max_Q \quad H(X \mid Y)
$$

subject to

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

worker constraints

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

item constraints

reference label

true label

worker label
Ordinal minimax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_Q \max_P H(X | Y)$$

subject to

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

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

difference from multiclass

worker constraints

item constraints

reference label

true label

worker label
Explaining the ordinal constraints

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

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

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

item difficulty \[ \tau_j(c, k) = \sum_{s \geq 1} \sum_{\Delta, \nabla} \tau_{j,s}^{\Delta, \nabla} \mathbb{I}(c \Delta s, k \nabla s) \]

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**

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

**item constraints**

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{i} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] \approx 0, \ \forall j, s$$
Regularized minimax conditional entropy

Jointly estimate $P$ and $Q$ by

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

subject to

worker constraints

$$\sum_{c \Delta s} \sum_{k \Delta s} \sum_{j} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = \xi_{i8}$$

item constraints

$$\sum_{c \Delta s} \sum_{k \Delta s} \sum_{i} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = \zeta_{j8}$$
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: metrics

• Evaluation metrics
  – L0 error: \[ L_0 = \mathbb{I}(y \neq \hat{y}) \]
  – L1 error: \[ L_1 = |y - \hat{y}| \]
  – L2 error: \[ L_2 = (y - \hat{y})^2 \]
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 et al. 2013)
Web search data

**Search Results**

- Machine learning - Wikipedia, the free encyclopedia
- Machine Learning | Coursera
- Machine Learning | Stanford Online
- Machine learning | Define Machine learning at Dictionary.com

**Rating:**

- 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

<table>
<thead>
<tr>
<th>Method</th>
<th>L0 Error</th>
<th>L1 Error</th>
<th>L2 Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority vote</td>
<td>0.269</td>
<td>0.428</td>
<td>0.930</td>
</tr>
<tr>
<td>Dawid &amp; Skene</td>
<td>0.170</td>
<td>0.205</td>
<td>0.539</td>
</tr>
<tr>
<td>Latent trait</td>
<td>0.201</td>
<td>0.211</td>
<td>0.481</td>
</tr>
<tr>
<td>Entropy multiclass</td>
<td>0.111</td>
<td>0.131</td>
<td>0.419</td>
</tr>
<tr>
<td>Entropy ordinal</td>
<td>0.104</td>
<td>0.118</td>
<td>0.384</td>
</tr>
</tbody>
</table>
Probabilistic labels vs error rates
Price prediction data

<table>
<thead>
<tr>
<th>Price Range</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 – $50</td>
<td>1</td>
</tr>
<tr>
<td>$51 – $100</td>
<td>2</td>
</tr>
<tr>
<td>$101 – $250</td>
<td>3</td>
</tr>
<tr>
<td>$251 – $500</td>
<td>4</td>
</tr>
<tr>
<td>$501 – $1000</td>
<td>5</td>
</tr>
<tr>
<td>$1001 – $2000</td>
<td>6</td>
</tr>
<tr>
<td>$2001 – $5000</td>
<td>7</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Method</th>
<th>L0 Error</th>
<th>L1 Error</th>
<th>L2 Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority vote</td>
<td>0.675</td>
<td>1.125</td>
<td>1.605</td>
</tr>
<tr>
<td>Dawid &amp; Skene</td>
<td>0.650</td>
<td>1.050</td>
<td>1.517</td>
</tr>
<tr>
<td>Latent trait</td>
<td>0.688</td>
<td>1.063</td>
<td>1.504</td>
</tr>
<tr>
<td>Entropy multiclass</td>
<td>0.675</td>
<td>1.150</td>
<td>1.643</td>
</tr>
<tr>
<td>Entropy ordinal</td>
<td>0.613</td>
<td>0.975</td>
<td>1.492</td>
</tr>
</tbody>
</table>
Summary

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