Methods for Ordinal Peer Grading

Karthik Raman, Thorsten Joachims

Talk: Wednesday 11:00am (Empire West)
Evaluation at Scale is challenging

• Conventional Evaluation:
  • Small-scale classes (10-15 students) : Instructors evaluate students themselves
  • Medium-scale classes (20-200 students) : TAs take over grading process.
  • MOOCs (10000+ students) : ??

★ MCQs & Other Auto-graded questions are not a good test of understanding.
★ Limits kinds of courses offered.
Peer Grading to the Rescue

- Students grade each other (anonymously)!
- Overcomes limitations of instructor/TA evaluation:
  - Number of “graders” scales with number of students!

- Current methods [Piech et. al. 13] require cardinal labels for each assignment.
- Each peer grader $g$ provides cardinal score for every assignment $d$ they grade.
  - E.g.: Likert Scale, Letter grade
Our Approach: Ordinal Peer Grading

• Challenge: Students are not trained graders.
  • Need to make feedback process simple!

• *Ordinal feedback* easier to provide and more reliable than *cardinal feedback*:
  • Project X is **better** than Project Y **vs.** Project X is a *B+.*

• *Ordinal Peer Grading*: Graders provide ordering of assignments they grade
  • Need to infer overall ordering and grader reliabilities.
Mallows Model and Variants

• GENERATIVE MODEL:

\[
P(\sigma^{(g)}|\sigma^*) = \frac{e^{-\delta_K(\sigma^*,\sigma^{(g)})}}{\sum_{\sigma'} e^{-\delta_K(\sigma^*,\sigma')}}
\]

\(\delta_K(\sigma^*,\sigma^{(g)})\) is the Kendall-Tau distance between orderings (# of differing pairs).

• OPTIMIZATION: NP-hard. Greedy algorithm provides good approximation.

• WITH GRADER RELIABILITY:

\[
P(\sigma^{(g)}|\sigma^*) = \frac{e^{-\eta_g \delta_K(\sigma^*,\sigma^{(g)})}}{\sum_{\sigma'} e^{-\eta_g \delta_K(\sigma^*,\sigma')}}
\]

• Variant with score-weighted objective (MALS) also studied.
Bradley-Terry Model & Variants

• **GENERATIVE MODEL:**
  
  \[ P(\sigma^{(g)} | s) = \prod_{d_i \succ_{\sigma^{(g)}} d_j} \frac{1}{1 + e^{-(s_{d_i} - s_{d_j})}} \]

  • Decomposes as pairwise preferences using logistic distribution of (true) score differences.

• **OPTIMIZATION:** Alternating minimization to compute MLE scores (and grader reliabilities) using SGD subroutine.

• **GRADER RELIABILITY:**
  
  \[ P(\sigma^{(g)} | s) = \prod_{d_i \succ_{\sigma^{(g)}} d_j} \frac{1}{1 + e^{-\eta_g(s_{d_i} - s_{d_j})}} \]

• Variants studied include Plackett-Luce model (PL) and Thurstone model (THUR).
Experimental Setting: New Peer Grading Dataset

- Data collected during class project (Fall 2013):
  - First real *large-scale* scale evaluation of machine-learning based peer-grading techniques.

- Used two-stages: Project Posters (PO) and Final-Reports (FR)
  - Students provided cardinal grades (10-point scale): 10-Perfect, 8-Good, 5-Borderline, 3-Deficient

- Also performed conventional grading: **TA and instructor grades**.

<table>
<thead>
<tr>
<th>Data Statistic</th>
<th>PO</th>
<th>FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Assignments</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>Number of Peer Reviewers</td>
<td>148</td>
<td>153</td>
</tr>
<tr>
<td>Total Peer Reviews</td>
<td>996</td>
<td>586</td>
</tr>
<tr>
<td>Total TA Reviews</td>
<td>78</td>
<td>88</td>
</tr>
<tr>
<td>Participating TAs</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>
How well do OPG methods do w.r.t. Instructor Grades?

- TAs had (Kendall-Tau) error of $22.0 \pm 16.0$ (Posters) and $22.2 \pm 6.8$ (Report).
Benefit of grader reliability: Identify poor graders

• Added lazy graders. Can we identify them?

• Significantly better than cardinal methods and simple heuristics.
• Survey shows most students found process valuable and feedback helpful.