Peer Prediction

conitzer@cs.duke.edu
Example setup

• We are evaluating a theme park which can be either Good or Bad
  • $P(G) = .8$

• If you visit, you can have an Enjoyable or an Unpleasant experience
  • $P(E|G) = .9$, $P(E|B) = .7$

• We ask people to report their experiences and want to reward them for accurate reporting

• The problem: *we will never find out the true quality / experience.*

• Another nice application: peer grading (of, say, essays) in a MOOC.
Solution: use multiple raters

- Rough idea: other agent likely (though not surely) had a similar experience
- Evaluate a rater by how well her report matches the other agent’s report
- How might this basic idea fail?

quality: good

I had fun.

I had fun.
Simple approach: output agreement

• Receive 1 if you agree, 0 otherwise

• What’s the problem?

• What is \( P(\text{other reports E} \mid \text{I experienced U}) \) (given that the other reports truthfully)?

\[
P(E' \mid U) = \frac{P(U \text{ and } E')}{P(U)}
\]

\[
P(U \text{ and } E') = P(U, E', G) + P(U, E', B) = .8 \cdot .9 + .2 \cdot .7 = .72 + .14 = .86
\]

\[
P(U) = P(U, G) + P(U, B) = .8 \cdot .9 + .2 \cdot .7 = .72 + .14 = .86
\]

So \( P(E' \mid U) = .114 / .14 = .814 \)

\[
P(E' \mid E) = \frac{P(E \text{ and } E')}{P(E)}
\]

\[
P(E \text{ and } E') = P(E, E', G) + P(E, E', B) = .8 \cdot .9 + .2 \cdot .7 = .72 + .14 = .86
\]

\[
P(E) = P(E, G) + P(E, B) = .8 \cdot .9 + .2 \cdot .7 = .72 + .14 = .86
\]

So \( P(E' \mid E) = .746 / .86 = .867 \)
The “1/Prior” mechanism [Jurca&Faltings’08]

• Receive $1/P(s)$ if you agree on signal $s$, 0 otherwise
• $P(E) = .86$ and $P(U) = .14$ so $1/P(E)=1.163$ and $1/P(U)=7.143$
• $P(E'|U) (1/P(E')) = .814*1.163 = .95$
• ... but, $P(U'|U)(1/P(U')) = .186*7.143=1.33$

• Why does this work? (When does this work?)
• Need, for all signals $s$, $t$: $P(s'|s)/P(s') > P(t'|s)/P(t')$
• Equivalently, for all signals $s$, $t$: $P(s,s')/P(s') > P(s,t')/P(t')$
• Equivalently, for all signals $s$, $t$: $P(s|s') > P(s|t')$
An example where the “1/Prior” mechanism does not work

• $P(A|\text{Good}) = .9$, $P(B|\text{Good}) = .1$, $P(C|\text{Good}) = 0$
• $P(A|\text{Bad}) = .4$, $P(B|\text{Bad}) = .5$, $P(C|\text{Bad}) = .1$
• $P(\text{Good}) = P(\text{Bad}) = .5$

• Note that $P(B|B') < P(B|C')$, so the condition from the previous slide is violated

• Suppose I saw B and the other player reports honestly

  • $P(B'|B) = P(B'|\text{Good}|B) + P(B'|\text{Bad}|B) = P(B'|\text{Good})P(\text{Good}|B) + P(B'|\text{Bad})P(\text{Bad}|B) = .1*(1/6) + .5*(5/6) = 13/30$
  
  • $P(B') = 3/10$, so expected reward for reporting B is $130/90 = 13/9 = 1.44$

  • $P(C'|B) = P(C'|\text{Good}|B) + P(C'|\text{Bad}|B) = P(C'|\text{Good})P(\text{Good}|B) + P(C'|\text{Bad})P(\text{Bad}|B) = 0*(1/6) + .1*(5/6) = 1/12$

  • $P(C') = 1/20$, so expected reward for reporting C is $20/12 = 5/3 = 1.67$
Better idea: use proper scoring rules

- **Assuming** the other reports truthfully, can infer a conditional distribution over the other’s report given my report
- Reward me according to a proper scoring rule!
- Suppose we use the logarithmic rule
- Reporting E $\Leftrightarrow$ predicting the other reports E’ with $P(E'|E) = .867$
- Reporting U $\Leftrightarrow$ predicting the other reports E’ with $P(E'|U) = .814$
- E.g., if report E and the other reports U’, I get $\ln(P(U'|E)) = \ln .133$
- In what sense does this work?
- Truthful reporting is an **equilibrium**
... as a Bayesian game

- A player’s type (private information): experience the player truly had (E or U)
- Note types are **correlated**
- (only displaying player 1’s payoffs)

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>true experiences:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E and E’ (prob. .746)</td>
<td>ln .867</td>
<td>ln .133</td>
</tr>
<tr>
<td>E and U’ (prob. .114)</td>
<td>ln .814</td>
<td>ln .186</td>
</tr>
<tr>
<td>U and E’ (prob. .114)</td>
<td>ln .867</td>
<td>ln .133</td>
</tr>
<tr>
<td>U and U’ (prob. .026)</td>
<td>ln .814</td>
<td>ln .186</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>E'</th>
<th>U'</th>
</tr>
</thead>
<tbody>
<tr>
<td>true experiences:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E and E’ (prob. .746)</td>
<td>ln .867</td>
<td>ln .133</td>
</tr>
<tr>
<td>E and U’ (prob. .114)</td>
<td>ln .814</td>
<td>ln .186</td>
</tr>
<tr>
<td>U and E’ (prob. .114)</td>
<td>ln .867</td>
<td>ln .133</td>
</tr>
<tr>
<td>U and U’ (prob. .026)</td>
<td>ln .814</td>
<td>ln .186</td>
</tr>
<tr>
<td>true experiences:</td>
<td>E'</td>
<td>U'</td>
</tr>
<tr>
<td>------------------</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>E and E' (prob. .746)</td>
<td>-0.143</td>
<td>-2.017</td>
</tr>
<tr>
<td>U and E' (prob. .114)</td>
<td>-0.205</td>
<td>-1.682</td>
</tr>
<tr>
<td>U and U' (prob. .026)</td>
<td>-0.143</td>
<td>-2.017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>true experiences:</th>
<th>E'</th>
<th>U'</th>
</tr>
</thead>
<tbody>
<tr>
<td>E and U' (prob. .114)</td>
<td>-0.143</td>
<td>-2.017</td>
</tr>
<tr>
<td>U and U' (prob. .026)</td>
<td>-0.205</td>
<td>-1.682</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>observe E: report E</td>
<td>observe U: report U</td>
<td>-0.406, -0.406</td>
<td>-0.154, -0.408</td>
</tr>
<tr>
<td>observe E: report E</td>
<td>observe U: report E</td>
<td>-0.408, -0.154</td>
<td>-0.143, -0.143</td>
</tr>
<tr>
<td>observe E: report U</td>
<td>observe U: report U</td>
<td>-0.415, -2.006</td>
<td>-0.205, -2.017</td>
</tr>
<tr>
<td>observe E: report U</td>
<td>observe U: report E</td>
<td>-0.417, -1.754</td>
<td>-0.200, -1.791</td>
</tr>
<tr>
<td>observe E: report U</td>
<td>observe U: report U</td>
<td>-0.415, -2.006</td>
<td>-0.205, -2.017</td>
</tr>
<tr>
<td>observe E: report U</td>
<td>observe U: report E</td>
<td>-0.417, -1.754</td>
<td>-0.200, -1.791</td>
</tr>
</tbody>
</table>
Downsides (and how to fix them, maybe?)

- **Multiplicity of equilibria**
  - Completely *uninformative* equilibria
  - Uselessly informative equilibria: Users may be supposed to evaluate whether the image contains a person, but instead reach an equilibrium where they evaluate whether the top-left pixel is blue

- Need to know the *prior distribution* beforehand
- Explicitly report beliefs as well [Prelec’04]
- **Bonus-penalty mechanism** [Dasgupta&Ghosh’13, Shnayder et al.’16]:
  - Suppose there are 3 tasks (e.g., 3 essays to grade)
  - You get a bonus for agreeing on the third task
    - Agents don’t know how the tasks are ordered
  - You get a penalty for agent 1’s report on 1 agreeing with agent 2’s report on 2
- Use a limited number of *trusted reports* (e.g., the instructor grades)
- ...?