Task 1: Predicting Citations

Claudia Perlich, Foster Provost & Sofus Macskassy

New York University

KDD Cup, August 2003

Claudia Perlich
Step 1: Feature Construction

- Temporal “shape” of citation counts:
  + counts for last 6 quarters and number of missing
- Seasonality of Publishing (conferences):
  + quarterly dummies
- Age of paper:
  + publication quarter
- Author reputation:
  + number of papers, total number of citations
  + median change at same paper “age”
- Paper reputation:
  + total and average number of citations per quarter

Claudia Perlich

---

Step 2: First Impression

- Very hard problem
  - Across quarters it is very hard to beat consistently a constant (Median -2 scores 1403)
- Distribution of target is not symmetric
  - L1 vs. L2 matters
- Strong but instable seasonality
  - The median delta varies
- Scaling differences across papers
  - Max citations: 2400, Average: 15
  - Only 10% of papers ever had more than 5 in a quarter

Claudia Perlich
Step 3: Training Considerations

- **Normalization**
  - Large variation in average citation counts
  - Normalize by average of the last 3 quarterly counts

- **Selection of Training Period**
  - Potentially non-stationary due to conference schedules, age of database, ...
  - Keep entire dataset

- **Selection of Training Observations**
  - Evaluation only on papers with at least 6 citations
  - Only use observation where the last citation count was at least 6

Claudia Perlich

---

Step 4: Model Class Selection

<table>
<thead>
<tr>
<th>Issues</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 for Evaluation</td>
<td>• L2 as cost function is wrong</td>
</tr>
<tr>
<td></td>
<td>• Median rather than mean</td>
</tr>
<tr>
<td>Integer Values</td>
<td>?</td>
</tr>
<tr>
<td>Noisy Task</td>
<td>Better bias than variance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Considered Options</th>
<th>L1</th>
<th>Integer</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>-</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>Neural Network</td>
<td>-</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>Ordered Probit</td>
<td>?</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Claudia Perlich
Ordered Probit Model

Results and Comparison

- Final Model:
  - 10 categories: -7, ..., +2
  - Ordered Probit with special prediction
  - Curtail predictions between -4 and 0

- Results:
  - Winning Participant: 1329
  - Best constant (-2): 1403
  - Curtailing Predictions between -4 and 0: 1360