Using Gradient Descent to Optimize Language Modeling Smoothing Parameters

Donald Metzler
University of Massachusetts Amherst

Supervised Parameter Estimation

**Direct Search**

Directly maximize evaluation metric by performing brute force search over parameter space.

**Pros:**
- No metric divergence
- Guaranteed to find maximum
- Works with any metric

**Cons:**
- Slow, especially with many parameters

**RankNet Cost Function**

The RankNet cost function is given by:

\[ C(Q, R) = \sum_{Q \in \mathcal{Q}} \sum_{(d_1, d_2) \in \mathcal{R}_Q} \log(1 + \exp(Y)) \]

where \( Y = g(Q; d_1) - g(Q; d_2) \) and \( \mathcal{R}_Q \) is the training data, such that if \((d_1, d_2)\) is in \( \mathcal{R}_Q \), then \( d_1 \) should be ranked higher than \( d_2 \). To minimize \( C \), we must compute:

\[
\frac{\delta C}{\delta \alpha} = \sum_{Q \in \mathcal{Q}} \sum_{(d_1, d_2) \in \mathcal{R}_Q} \frac{\delta C}{\delta Y} \frac{\delta Y}{\delta \alpha}
\]

\[
\frac{\delta C}{\delta \beta} = \frac{\exp[g(Q; d_1) - g(Q; d_2)]}{1 + \exp[g(Q; d_1) - g(Q; d_2)]} \frac{\delta g(Q; d_1)}{\delta \alpha} - \frac{\delta g(Q; d_1)}{\delta \beta}
\]

**Pros:**
- Efficient
- Easy to optimize (for differentiable \( g(Q; D) \))
- Global optimum (for convex \( g(Q; D) \))

**Cons:**
- Does not necessarily work well with all metrics

Language Modeling

**Two-Stage Language Model**

- Proposed by Zhai and Lafferty in 2002
- Robust language modeling estimate
- Combines Jelinek-Mercer and Dirichlet smoothing
- Two smoothing parameters must be estimated
- Unsupervised estimation possible

**Scoring Function**

Documents ranked according to the log of the query likelihood. That is,

\[
g(Q; D) = \sum_{w \in Q} \log \left( \frac{1 - \lambda}{1 - \mu} \frac{f_{w,D} + \mu P(w|C)}{P(D)} + \lambda P(w|C) \right)
\]

**Partial Derivatives**

In order to optimize the RankNet cost function, partial derivatives of the scoring function must be computed. These derivatives are:

\[
\frac{\delta g(Q; D)}{\delta \lambda} = \sum_{w \in Q} \frac{\delta}{\delta \lambda} \left( \frac{1 - \lambda}{1 - \mu} \frac{P(w|C)}{P(D)} + \lambda P(w|C) \right)
\]

\[
\frac{\delta g(Q; D)}{\delta \mu} = \sum_{w \in Q} \frac{1}{1 - \mu} \frac{\delta}{\delta \mu} \left( \frac{P(w|C)}{P(D)} - \frac{1 - \lambda - \lambda P(w|C)}{1 - \lambda} \right)
\]

Gradient descent can be used to find the setting of smoothing parameters that minimizes the RankNet cost function.

Evaluation

**Setup**

- Four TREC data sets used
- Topics split into training and test sets
- Training data consists of all pairwise preferences

**Results**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Estimate</th>
<th>ap</th>
<th>wjg</th>
<th>robust</th>
<th>wt10g</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>Direct</td>
<td>.2572</td>
<td>.2555</td>
<td>.2920</td>
<td>.1930</td>
</tr>
<tr>
<td></td>
<td>Optimal</td>
<td>.2794</td>
<td>.2766</td>
<td>.2756</td>
<td>.1923</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.2668</td>
<td>.2631</td>
<td>.2620</td>
<td>.1999</td>
</tr>
<tr>
<td>BPREF</td>
<td>Direct</td>
<td>.3470</td>
<td>.3392</td>
<td>.2821</td>
<td>.1834</td>
</tr>
<tr>
<td></td>
<td>Optimal</td>
<td>.3487</td>
<td>.3394</td>
<td>.2801</td>
<td>.1845</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.3557</td>
<td>.3454</td>
<td>.2849</td>
<td>.1907</td>
</tr>
<tr>
<td>P910</td>
<td>Direct</td>
<td>.4840</td>
<td>.4850</td>
<td>.4323</td>
<td>.3294</td>
</tr>
<tr>
<td></td>
<td>RankNet</td>
<td>.4840</td>
<td>.4834</td>
<td>.4384</td>
<td>.3143</td>
</tr>
<tr>
<td></td>
<td>Optimal</td>
<td>.4960</td>
<td>.4960</td>
<td>.4434</td>
<td>.3388</td>
</tr>
</tbody>
</table>

Table 1: Test set effectiveness for parameters estimated using direct search and RankNet. Optimal effectiveness values are also provided as an upper bound. Effectiveness is measured in terms of mean average precision, binary precision, and precision at 10. Italicized values indicate statistically significant improvements over direct search. Bold values indicate significant improvements over RankNet.

**Conclusions**

- RankNet is never significantly better than Direct
- Direct sometimes significantly better than RankNet
- RankNet generalizes similarly for all metrics
- Direct generalizes better across metrics