



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:

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

where $Y = g(Q; d_2) - g(Q; d_1)$ 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:

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

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

$$\frac{\delta Y}{\delta \alpha} = \frac{\delta g(Q; d_2)}{\delta \alpha} - \frac{\delta g(Q; d_1)}{\delta \alpha}$$

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,

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

Partial Derivatives

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

$$\frac{\delta g(Q; D)}{\delta \lambda} = \sum_{w \in Q} \frac{P(w|C) - \frac{tf_{w,D} + \mu P(w|C)}{\mu + |D|}}{(1 - \lambda) \frac{tf_{w,D} + \mu P(w|C)}{\mu + |D|} + \lambda P(w|C)}$$

$$\frac{\delta g(Q; D)}{\delta \mu} = \sum_{w \in Q} \frac{(1 - \lambda) \frac{|D|}{(\mu + |D|)^2} \left(P(w|C) - \frac{tf_{w,D}}{|D|} \right)}{(1 - \lambda) \frac{tf_{w,D} + \mu P(w|C)}{\mu + |D|} + \lambda P(w|C)}$$

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

Metric	Estimate	ap	wsj	robust	wt10g
MAP	Direct	.2072	.3255	.2920	.1930
	RankNet	.2081	.2987	.2756	.1922
	Optimal	.2088	.3290	.2931	.2003
BPREF	Direct	.3490	.3392	.2821	.1834
	RankNet	.3437	.3294	.2661	.1805
	Optimal	.3504	.3557	.2840	.1907
P@10	Direct	.3360	.4820	.4323	.3204
	RankNet	.3400	.4240	.4384	.3143
	Optimal	<i>.3520</i>	.4960	.4434	.3388

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 preference, 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