Modeling Label Ambiguity for List-Wise Neural Learning to Rank

Jagerman, R.; Kiseleva, J.; de Rijke, M.

Publication date
2017
Document Version
Author accepted manuscript
Published in
Neu-IR: Workshop on Neural Information Retrieval

Modeling Label Ambiguity for Neural List-Wise Learning to Rank

Rolf Jagerman
University of Amsterdam
Amsterdam, The Netherlands
rolf.jagerman@uva.nl

Julia Kiseleva
UserSat.com & University of Amsterdam
Amsterdam, The Netherlands
j.kiseleva@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

ABSTRACT

List-wise learning to rank methods are considered to be the state-of-the-art. One of the major problems with these methods is that the ambiguous nature of relevance labels in learning to rank data is ignored. Ambiguity of relevance labels refers to the phenomenon that multiple documents may be assigned the same relevance label for a given query, so that no preference order should be learned for those documents. In this paper we propose a novel sampling technique for computing a list-wise loss that can take into account this ambiguity. We show the effectiveness of the proposed method by training a 3-layer deep neural network. We compare our new loss function to two strong baselines: ListNet and ListMLE. We show that our method generalizes better and significantly outperforms other methods on the validation and test sets.

CCS CONCEPTS

•Computing methodologies →Neural networks; •Information systems →Learning to rank;

KEYWORDS

Neural Information Retrieval; List-wise Learning to Rank

1 INTRODUCTION

One of the most important components to any search engine is the Learning to Rank (LTR) model. It considers dozens or even hundreds of relevance signals and determines in what order to show the documents to the user based on these signals. The following three main directions have emerged in the field of LTR:

(1) **Point-wise** [4–6]: Models the LTR problem as a (usually probabilistic) regression problem.
(2) **Pair-wise** [1, 8]: Casts the LTR problem as a classification problem and learns preferences between pairs of documents.
(3) **List-wise** [2, 3, 15]: Attempts to solve LTR by treating lists of preferences as instances for learning; these methods are considered to be the current state-of-the-art.

In this paper, we focus on list-wise LTR, since these methods are the current state-of-the-art and are commonly used in conjunction with neural networks. In particular, we focus on ListNet [3] and ListMLE [15]. One of the major difficulties with these list-wise methods is that there is no consideration for the ambiguity that exists in LTR data that uses relevance scores.

In LTR we deal with relevance labels that score documents on a finite ordinal scale. Let us consider an example in Figure 1.

Figure 1: Relevance labels in LTR admit many different "correct" rankings for the same query q. The colors indicate different relevance grades: green is highly relevant, blue is relevant and red is not relevant.

There are typically more documents than relevance labels, which necessarily introduces ambiguity in the rankings. Any documents that share the same relevance label can freely be interchanged, thus any permutation of documents with the same relevance label is technically correct. As a consequence, there are many possible rankings of the documents that would be considered "correct" for the same query q. That brings us to the problem of *label ambiguity* which is the phenomenon that multiple documents may be assigned the same relevance label for a given query, so that no preference order should be learned for those documents. Learning a preference where none exists may lead to overfitting or limitations in the learner's ability to generalize.

In ListNet, it is computationally too expensive to model label ambiguity, because every possible permutation has to be considered. Cao et al. [3] address this problem by introducing a top-k approximation. However, we argue that this largely mitigates the major attractiveness of ListNet, namely its capability to learn from the full ranked list. For ListMLE, Xia et al. [15] make a simplifying

SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17), August 7–11, 2017, Shinjuku, Tokyo, Japan

DOI: 10.1145/nnnnnnn.nnnnnnn

assumption by sampling one perfect ranking and assuming that this is the ground truth. These drawbacks lead to our main research question:

> *Can we learn a list-wise neural model while taking into account the ambiguity of relevance labels in LTR data?*

Our contributions are two-fold:

- We introduce a novel way to optimize a list-wise neural LTR model, sampling learning instances directly from the Plackett-Luce distribution [10, 12], to take into account the ambiguity of relevance labels in an LTR setting.
- We publish the source code of our method, which uses Chainer [14], a GPU-accelerated deep learning framework, promoting future research in neural list-wise LTR methods.[1]

The remainder of this paper is organized as follows. Section 2 introduces our notation and earlier work on list-wise LTR approaches. Section 3 discusses in details the phenomena of label ambiguity. We present our solution in Section 4. We present our experimental results in Section 5. Finally, we conclude in Section 6.

2 RELATED WORK

In this section we discuss several important works that our method builds on. First, we introduce some notation that is used throughout this paper (Section 2.1). Next, we briefly discuss the Plackett-Luce (PL) distribution (Section 2.2), followed by ListNet [3] (Section 2.3) and ListMLE [15] (Section 2.4).

2.1 Preliminaries

We consider the scenario of Learning to Rank (LTR) where we are given a collection of m queries $\mathbb{Q} = \{q^{(1)}, \ldots, q^{(m)}\}$. Each query $q^{(i)}$ is associated with a set of n documents $\mathbb{D}^{(i)} = \{\boldsymbol{d}_1^{(i)}, \ldots, \boldsymbol{d}_n^{(i)}\}$ and corresponding relevance labels $\mathbb{Y}^{(i)} = \{y_1^{(i)}, \ldots, y_n^{(i)}\}$. Each document $\boldsymbol{d}_j^{(i)}$ is a d-dimensional feature vector representing the query-document pair. For the sake of brevity, we will drop the superscript notation $\cdot^{(i)}$ for the remainder of this paper.

In LTR we use a scoring function f to score every document $\boldsymbol{d}_j$ and then sort the set of documents by these scores. Our objective is to find a function f for which the resulting ranking is optimal with regards to the relevance labels. In other words, we wish to find a function f that assigns high scores to documents that have high relevance and low scores to documents that have little relevance.

2.2 Plackett-Luce Distribution

The Plackett-Luce (PL) distribution [10, 12] is used extensively in probabilistic list-wise LTR methods. It is based on the idea that a ranking is drawn sequentially from a list of item-specific scores, one item at a time. The PL distribution is a probability distribution over all possible permutations of a set of item scores. Intuitively, it assigns a high probability to permutations that place high-scoring items at the top while assigning low probability to permutations that place high-scoring items at the bottom. More formally, the PL probability for a permutation $\pi = \{\pi_1, \ldots, \pi_n\}$, given a scoring function f and a set of documents $\mathbb{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n\}$ is defined as follows:

$$PL(\pi \mid \mathbb{D}; f) = \prod_{i=1}^{n} \frac{\phi(f(\boldsymbol{d}_{\pi_i}))}{\sum_{j=i}^{n} \phi(f(\boldsymbol{d}_{\pi_j}))}, \tag{1}$$

where $\phi : \mathbb{R} \rightarrow \mathbb{R}_{>0}$ is a function that maps real-valued scores to positive numbers. In practice, this is usually chosen to be the exponential function. For the sake of simplifying the derivations in Section 4, we also choose the exponential function. Hence, we will from now on assume $\phi(\cdot) = \exp(\cdot)$.

We consider two LTR methods that build on the PL distribution: ListNet [3] and ListMLE [15]. It can be shown that the losses associated with these methods are continuous, convex and differentiable [15]. This makes them easy to optimize via Stochastic Gradient Descent (SGD) and attractive in a neural setting.

2.3 ListNet

ListNet [3] is one of the earliest list-wise LTR methods. It considers the following two PL distributions:

- $PL(\pi \mid \mathbb{D}; f)$
  This is the PL distribution of the output of the network f. Any permutation π that places documents that generated large network scores at the top gets assigned high probability.
- $PL(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}})$
  This is the PL distribution of a mapping of the relevance labels $\mathbb{Y}$. The mapping $\psi_{\mathbb{Y}}$ generates a score vector of the ground truth that retains the order of the relevance labels:

  $$\psi_{\mathbb{Y}}(y_i) > \psi_{\mathbb{Y}}(y_j) \iff y_i > y_j \tag{2}$$

  Any permutation π that places documents with high relevance at the top gets assigned high probability.

The optimization objective then becomes minimizing the Kullback-Leibler divergence of these two distributions:

$$\min_f D_{\mathrm{KL}}\left(PL(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}) \;||\; PL(\pi \mid \mathbb{D}; f)\right) \tag{3}$$

This is equivalent to minimizing the cross entropy, giving rise to the following loss function:

$$\mathcal{L}(f(\mathbb{D}), \mathbb{Y}) = -\sum_{\pi \in \Omega} PL(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}) \log\left(PL(\pi \mid \mathbb{D}; f)\right), \tag{4}$$

where Ω is the set of all possible permutations.

To compute this cross entropy loss, we have to consider every possible permutation in Ω, which is of the size $O(n!)$. Cao et al. [3] resort to a top-k approximation, where k is usually 1, to make the computation feasible.

A stochastic top-k ListNet variant has been proposed by Luo et al. [11]. In this paper, the authors sample from within the top-k subgroups to speed up training. This is different to our work, where we completely eliminate the need for a top-k approximation.

2.4 ListMLE

ListMLE [15] replaces the optimization objective of ListNet with a simpler form. The simplifying assumption is that a single permutation $\pi \in \{\pi \mid y_{\pi_i} \geq y_{\pi_j}; i < j\}$ is chosen and considered to be *the ground truth*. It then directly optimizes the PL probability of

[1] https://github.com/rjagerman/shoelace

the network scores for this permutation and uses the negative log probability as a loss:

$$\mathcal{L}\left(f(\mathbb{D}), \mathbb{Y}\right) = -\log PL\left(\pi \mid \mathbb{D}; f\right). \tag{5}$$

One of the main drawbacks of ListMLE is that it assumes that a single perfect ranking π is known. This assumption however does not hold for LTR data sets, where we have ambiguous relevance labels.

To summarize, both ListNet and ListMLE build on the PL distribution and provide elegant probabilistic ways to learn a list-wise LTR model. The key distinction of our work compared to previous efforts is that we introduce a new LTR method that properly deals with the *label ambiguity* problem, which we will describe next.

3 THE AMBIGUITY OF RANKING

Existing work on list-wise approaches typically ignores the ambiguity of the labels. For instance, ListMLE samples a *single* permutation from the ground truth, whose ordering is then assumed to be *the* ground truth ranked list. For example, take the following 8 documents with corresponding relevance scores:

$$\begin{array}{rcccccccc} \mathbb{D} = [& \boldsymbol{d}_1 & \boldsymbol{d}_2 & \boldsymbol{d}_3 & \boldsymbol{d}_4 & \boldsymbol{d}_5 & \boldsymbol{d}_6 & \boldsymbol{d}_7 & \boldsymbol{d}_8 &] \\ \mathbb{Y} = [& 2 & 2 & 1 & 1 & 1 & 0 & 0 & 0 &] \end{array} \tag{6}$$

What we wish to learn are the following preferences:

$$\{\boldsymbol{d}_1 \leftrightarrow \boldsymbol{d}_2\} \succ \{\boldsymbol{d}_3 \leftrightarrow \boldsymbol{d}_4 \leftrightarrow \boldsymbol{d}_5\} \succ \{\boldsymbol{d}_6 \leftrightarrow \boldsymbol{d}_7 \leftrightarrow \boldsymbol{d}_8\}. \tag{7}$$

Instead, the optimization objective in ListMLE learns the following:

$$\boldsymbol{d}_1 \succ \boldsymbol{d}_2 \succ \boldsymbol{d}_3 \succ \boldsymbol{d}_4 \succ \boldsymbol{d}_5 \succ \boldsymbol{d}_6 \succ \boldsymbol{d}_7 \succ \boldsymbol{d}_8. \tag{8}$$

Thus, the learning algorithm will attempt to learn $\boldsymbol{d}_1 \succ \boldsymbol{d}_2$. This is problematic, because according to the ground truth, the relative ordering of $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ is not meaningful. Attempting to learn such overly specific relations is harmful to the generalization power of the learning algorithm.

Next, we will present our method that overcomes the described issue.

4 SAMPLING RANKINGS FROM THE PLACKETT-LUCE DISTRIBUTION

Instead of naively choosing a single permutation of the documents π and considering that permutation to be the ground truth, we propose a more sophisticated sampling method. The main idea is to directly sample a ranking from the PL distribution of the relevance labels during every stochastic update.

To motivate this decision from a theoretical point of view, let us revisit the ListNet cross entropy loss function. We can rewrite Equation 4 into the following form:

$$\mathcal{L}\left(f(\mathbb{D}), \mathbb{Y}\right) = \sum_{\pi \in \Omega} \underbrace{PL\left(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}\right)}_{\text{weight}} \underbrace{\left(-\log\left(PL\left(\pi \mid \mathbb{D}; f\right)\right)\right)}_{\text{loss}} \tag{9}$$

This loss can be interpreted as a stochastic variant of the ListMLE loss. Here we sample a possible permutation π with a corresponding probability $PL\left(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}\right)$. We can then use that sample to compute a stochastic loss. We call this method ListPL:

$$\begin{aligned} \mathcal{L}\left(f(\mathbb{D}), \mathbb{Y}\right) &= -\log\left(PL\left(\pi \mid \mathbb{D}; f\right)\right) \\ \pi &\sim PL\left(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}\right) \end{aligned} \tag{10}$$

Deriving the stochastic gradient of this loss function follows the same derivation as ListMLE. We include the derivative with respect to the activation function here for completeness sake:

$$\begin{aligned} &\frac{\partial}{\partial f(\boldsymbol{d}_{\pi_k})}\left[-\log\left(PL\left(\pi \mid \mathbb{D}; f\right)\right)\right] \\ &= -\frac{\partial}{\partial f(\boldsymbol{d}_{\pi_k})}\left[\log \prod_{i=1}^{n} \frac{\exp(f(\boldsymbol{d}_{\pi_i}))}{\sum_{j=i}^{n} \exp(f(\boldsymbol{d}_{\pi_j}))}\right] \\ &= -\frac{\partial}{\partial f(\boldsymbol{d}_{\pi_k})}\left[\sum_{i=1}^{n}\left(f(\boldsymbol{d}_{\pi_i}) - \log \sum_{j=i}^{n} \exp(f(\boldsymbol{d}_{\pi_j}))\right)\right] \\ &= \sum_{i=1}^{n} \mathbf{1}_{i \le k \le n}\left(\frac{\exp(f(\boldsymbol{d}_{\pi_k}))}{\sum_{j=i}^{n} \exp(f(\boldsymbol{d}_{\pi_j}))}\right) - 1 \\ &= \sum_{i=k}^{n}\left(\frac{\exp(f(\boldsymbol{d}_{\pi_k}))}{\sum_{j=i}^{n} \exp(f(\boldsymbol{d}_{\pi_j}))}\right) - 1 \end{aligned} \tag{11}$$

The resulting loss and corresponding gradient can then be used to train a neural network via SGD.

A different way to look at ListPL is like a data set generation method. We are essentially applying ListMLE to a data set that contains all permutations π of possible rankings and weighing each one by $PL\left(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}\right)$. Looking at the method from this perspective, it is reasonable to assume we can do sampling-based SGD the usual way: sample a ranking π uniformly and then weigh the gradient by $PL\left(\pi \mid \mathbb{D}; \psi_{\mathbb{Y}}\right)$. However, this runs into a problem: the sample space is enormous ($O(n!)$). With sufficiently many documents per query, most uniformly sampled rankings will have a PL probability that is close to 0 resulting in a very slow convergence rate.

In this section, we described our method, ListPL, which is able to deal with the label ambiguity problem described in Section 3. Next, we will present our experimental evaluation.

5 EXPERIMENTS AND RESULTS

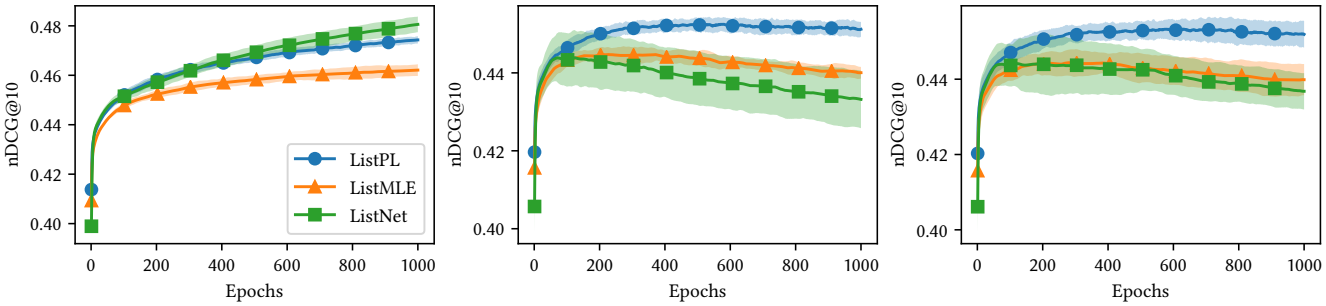
To validate the effectiveness of our method, ListPL, we compare it to ListNet (top-1) and ListMLE on the MSLR-WEB10k data set [13]. This data set contains 10,000 queries. It represents query-document pairs as 136-dimensional feature vectors and grades them on a scale from 0 (irrelevant) to 4 (perfectly relevant). We use 6,000 queries for training, 2,000 for validation and 2,000 for testing.

The architecture of our network is a 3-layer fully connected neural network with 80 ReLU activation units at each hidden layer. We experimented with more than 3 layers and found only negligible improvements. We keep the network architecture the same and vary only the loss function:

- ListNet loss using a top-1 approximation (Equation 4)
- ListMLE loss (Equation 5)
- ListPL loss (Equation 10)

ADAM [9] is used as the optimizer with the default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate $\alpha = 10^{-5}$. The experiments are run for 1000 epochs.

Figure 2 shows the results on the MSLR-WEB10k data set. We evaluate the performance of the methods using nDCG@10 [7], which is a natural evaluation metric for LTR in the Web-search setting. We see that ListPL performs similar to ListNet during training,



(a) Training set (b) Validation set (c) Test set

Figure 2: nDCG@10 performance on MSLR-WEB10K. The shaded areas indicate the standard deviation over the different folds. On the test set, the performance improvements over ListNet and ListMLE are statistically significant with $p = 0.00078$ and $p = 0.00218$ respectively (two-tailed t-test).

but outperforms both ListNet and ListMLE during validation and testing. The performance degradations on the validation set and test set for ListNet and ListMLE that occur after 100 epochs indicate that these methods are overfitting and are effectively learning noise coming from label ambiguity. These results are in line with our expectations because ListPL properly deals with the ambiguity in the relevance scores and thus generalizes better.

The nDCG@10 performance on the test set is evaluated using 5-fold cross validation. The performance improvement of ListPL over ListNet is statistically significant with $p = 0.00078$ (two-tailed t-test) whereas the performance improvement over ListMLE is statistically significant with $p = 0.00218$ (two-tailed t-test).

To summarize, based on extensive experimentation with the MSLR-WEB10k data set, we conclude that ListPL significantly outperforms strong baselines due to the fact that it handles the label ambiguity problem well, and, hence, generalizes better.

6 CONCLUSION

The paper extends earlier work on list-wise approaches for LTR [2, 3, 15]. Specifically, we have considered the problem of label ambiguity. Our main research question was:

> *Can we learn a list-wise neural model while taking into account the ambiguity of relevance labels in LTR data?*

Our overall conclusion is that by introducing a sampling method based on directly sampling from the Plackett-Luce (PL) distribution of relevance labels, we are able to increase the ability to generalize neural list-wise LTR methods while maintaining efficiency. Specifically, we used a modified loss function that can efficiently mitigate the problem of label ambiguity and thereby improve over existing list-wise neural LTR methods.

Our extensive experimentation with the MSLR-WEB10k data set showed that our method, ListPL, significantly outperforms two strong baselines: ListNet [3] and ListMLE [15]. Our method and baselines are implemented using Chainer [14], a GPU-accelerated deep learning framework. We are sharing the source code of ListPL online[2] (MIT licensed), which we hope is useful for future research towards list-wise neural methods.

[2]https://github.com/rjagerman/shoelace

Acknowledgments. This research was supported by Ahold Delhaize, Amsterdam Data Science, the Bloomberg Research Grant program, the Criteo Faculty Research Award program, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Microsoft Research Ph.D. program, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.001.116, HOR-11-10, CI-14-25, 652.002.001, 612.001.551, 652.001.003, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *ICML*. ACM, 89–96.
[2] Christopher JC Burges, Robert Ragno, and Quoc Viet Le. 2006. Learning to rank with nonsmooth cost functions. In *NIPS*, Vol. 6. 193–200.
[3] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*. ACM, 129–136.
[4] William S Cooper, Fredric C Gey, and Daniel P Dabney. 1992. Probabilistic retrieval based on staged logistic regression. In *SIGIR*. ACM, 198–210.
[5] Koby Crammer, Yoram Singer, and others. 2001. Pranking with ranking. In *NIPS*, Vol. 14. 641–647.
[6] Norbert Fuhr. 1989. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems (TOIS)* 7, 3 (1989), 183–204.
[7] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *SIGIR*. ACM, 41–48.
[8] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD*. ACM, 133–142.
[9] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[10] R Duncan Luce. 2005. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation.
[11] Tianyi Luo, Dong Wang, Rong Liu, and Yiqiao Pan. 2015. Stochastic top-k ListNet. In *EMNLP*. 676–684.
[12] Robin L Plackett. 1975. The analysis of permutations. *Applied Statistics* 24, 2 (1975), 193–202.
[13] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *CoRR* abs/1306.2597 (2013). http://arxiv.org/abs/1306.2597
[14] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: A next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.
[15] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *ICML*. ACM, 1192–1199.