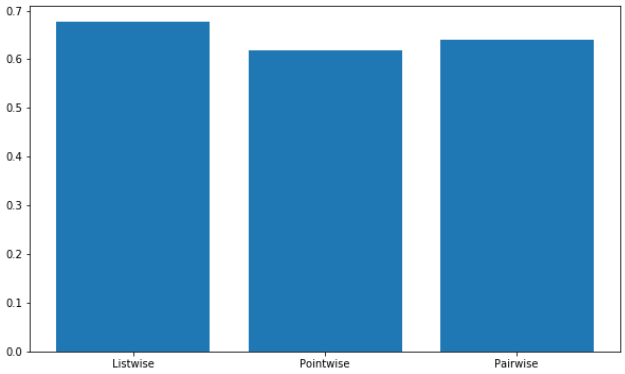
Report for Information Retrieval Assignment 3

Mircea Mironenco
mircea.mironenco@student.uva.nl

1 Analysis

All 3 algorithms have been implemented and tested. The overall NDCG results can be seen in figure 1. The neural network architecture used to produce the results is the default one, specifically with 64 input features, 200 hidden units, a learning rate of 0.00005 and momentum of 0.95.

Figure 1: Final averaged NDCG values for all algorithms



We can observe that using 5-fold cross-validation the overall best average is obtained by the LambdaRank algorithm, outperforming RankNet, and the worst by the Pointwise algorithm. While the network was trained for only 5 epochs, we can start to observe small differences in terms of performance between the algorithms.

The differences in performance can be motivated by the fact that pointwise will attempt to predict correctly all zero document labels. Moreover, the algorithm is likely biased to predict 1s as 0s due to the sparsity of the data.

Using RankNet we are able to push more relevant documents to the top and less relevant ones to the bottom. As opposed to the Pointwise algorithm, RankNet is optimizing a pairwise error, i.e. a relationship, using neural networks to create a latent space where it can differentiate between 2 ranks given their level of relevancy as a real number.

However, it does not incorporate Information Retrieval measures in it's calculations, which leaves room for improvement. LambdaRank incorporates the ordering of the list, by considering the exact ranking of each item through the NDCG measure.

The final results for all tests sets, considering all algorithms are presented in the tables below. Because we have only trained the algorithms for 5 epochs, the differences between the performance of the algorithms is still very small.

.

Table 1: All test set NDCG results

	Pointwise	RankNet	LambdaRank
Test #1	0.55	0.64	0.67
Test #2	0.73	0.66	0.69
Test #3	0.63	0.70	0.74
Test #4	0.66	0.62	0.64
Test #5	0.49	0.56	0.62

Overall averages presented in Figure 1 are:

- Pointwise: 0.617
- Pairwise: 0.64
- Listwise: 0.67

2 Conclusion

While the results tend to favor the LambdaRank algorithm, it is unclear if we were to replace the NDCG measure with another similar offline evaluation measure (such as ERR/RBP), that the performance would still be similar. Furthermore, the experiments were not run until convergence, due to time constraints, only 5 epochs were considered.