Homework 6

1 Introduction

1.1 Collaboration and Originality

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TAs.

   No

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

   No

3. Did you examine anyone else's software for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   No.

4. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   Yes.

5. Are you the author of every word of your report (Yes or No)?

   Yes

1.2 Instructions

2 Experiment: Baselines

2.1 Experimental Results

	BM25 (Exp-2.1a)	**Indri (Exp-2.1b)**	**LTR (Exp-2.1c)**	**DRMM (Exp-2.1d)**
P@10	0.4409	0.4136	0.3818	0.2682
P@20	0.4318	0.4341	0.425	0.2932
P@30	0.4273	0.4485	0.4273	0.3212
NDCG@10	0.3589	0.3149	0.279	0.2011
NDCG@20	0.3571	0.3397	0.3094	0.2104
NDCG@30	0.3592	0.3572	0.3257	0.2327
MAP	0.2603	0.2729	0.2542	0.1495
Time	00:27	00:33	01:11	05:41

2.2 Parameters

For BM25/Indri

BM25:k_1=1.2

BM25:b=0.75

BM25:k_3=0

Indri:mu=1500

Indri:lambda=0.01

For LetoR

One provied file for trainQry, one provide file for trainQrels

For DRMM:

drmm:numHistogramBins = 30

drmm:mlpLayers = 5,1

drmm:numEpochs = 2500

drmm:numTrainingPairs = 5000

drmm:learningRate = 0.01

drmm:randomSeed=11642

2.3 Discussion

The DRMM algorithm underperforms traditional sophisticated BM25, Indri and LTR algorithm. Moreover, it takes a lot of time to run. But note that, the time shown here contains both training and testing time (same as LTR). It is known that the Neural Network models take lots of data and lots of computation to generate good results. And the forward passing is less computationally intensive than the backward passing. Therefore, the testing time will be much smaller than 05:41 and will be less different from the traditional algorithm.

Another phenomenon is that both learning algorithms, LTR and DRMM do not perform better than the rule-based, feature-based traditional models. The reason could be the limit of training data, training time and/or limit model complexity.

3 Experiment: Number of Histogram Bins

3.1 Experimental results

	11 (Exp-3.1a)	31 (Exp-3.1b)	6 (Exp-3.1c)	51 (Exp-3.1d)
P@10	0.2182	0.2727	0.3364	0.2545
P@20	0.2818	0.2591	0.3136	0.2932
P@30	0.2924	0.2788	0.3242	0.3045
NDCG@10	0.1894	0.1999	0.2806	0.1844
NDCG@20	0.2154	0.1984	0.2603	0.215
NDCG@30	0.2236	0.2138	0.2626	0.2237
MAP	0.139	0.1248	0.1525	0.1262
Time	05:41	05:17	04:56	05:24

3.2 Parameters

drmm:mlpLayers=5,1

drmm:numEpochs=2500

drmm:numTrainingPairs=5000

drmm:learningRate=0.01

drmm:randomSeed=11642

3.3 Discussion

It does not mean the higher the bin number is the better the performance. And in my experiments, the bin size of 6 has the best performance. The higher the dimension in neural network could store more information. However, if the training data is not enough, having a high dimension will result in over parametrization. The model is much easier to overfit. The parameters of neural network could simply remember the training sample, instead of extracting some useful feature from them. Each dimension will be noisy and might not be meaningful. In our setting, the data is too sparse, therefore the lower dimension neural network can better extract features. The performance of bin 6 is much better than bin 11. However, it is not the case that bin 11 outperforms bin 15 and 31. And this is a counter example of previous analysis. And this sort of shows that the hyperparameter of neural network is hard to tune. It is also possible that splitting score into bins of different granularity results in very different data and some of them happen to be easy to train and others not.

The time spending on different size of bin is also different. But it is interesting that the computing time is not strictly increasing with the growth of bins. Bin 31 results in slightly faster computing than bin 11. This small different could be due to runtime environment, system reason. But bin 6 computes much faster than others.

4 Experiment: Effect of Training Effort

4.1 Experimental results

Vary Number of Epochs				
	0 **(Exp-4.1a)**	1 **(Exp-4.1b)**	1000 **(Exp-4.1c)**	10000 **(Exp-4.1d)**
P@10	0.2318	0.2409	0.3364	0.3045
P@20	0.2318	0.2341	0.3341	0.3227
P@30	0.253	0.2485	0.3197	0.3258
NDCG@10	0.1596	0.1582	0.2823	0.2444
NDCG@20	0.161	0.1634	0.2707	0.2441
NDCG@30	0.1766	0.1735	0.267	0.2504
MAP	0.1224	0.1192	0.1545	0.1532
Time				

Vary Number of Pairs				
	1 **(Exp-4.2a)**	2500 **(Exp-4.2b)**	2501 **(Exp-4.2c)**	10000 **(Exp-4.2d)**
P@10	0.2364	0.2909	0.3	0.3182
P@20	0.2455	0.3182	0.3159	0.3023
P@30	0.25	0.3152	0.3152	0.3136
NDCG@10	0.163	0.2174	0.218	0.2365
NDCG@20	0.1677	0.2386	0.2356	0.2263
NDCG@30	0.1748	0.2411	0.239	0.2347
MAP	0.1203	0.1461	0.1466	0.152
Time				

4.2 Parameters

drmm:mlpLayers=5,1

drmm:numTrainingPairs=5000

drmm:learningRate=0.01

drmm:randomSeed=11642

For the first group of experiments:

Pairs is fiexed to 2500.

For the second group of experiments:

Since the first exp has only 1 pair, it only trains for 1 epoch.

For the rest, they train for 5000 epochs.

4.3 Discussion

The effect of epochs.

The first experiment does not train at all, but it has MAP of 0.1224. Since the model is randomly initialized, we can consider this as a random guess. For training only one epoch, in our experiment, the performance does not change much. The P@10 increases but the overall MAP drops. This shows that the update of the neural network is noisy at each step but on overage it is improving. Therefore, if we look at the performance after training 1000 epochs, the result improves a lot. And the difference is the result of learning. Finally, if we compare training 1000 epochs with 10000 epochs, we can see there is no more improvement. Moreover, the performance decreases. This could be the result of over fitting the parameter to the training set. As we can see the Precision at the top range deteriorates a lot. The training epoch should be tuned according to the volume of training data.

The effect of pairs.

Overall, it seems the more training data, the result is better. And it is interesting to see that, by increasing one more pair, from 2500 to 2501. The overall performance of the model increases.

5 Experiment: The Effect of the Feedforward Network

5.1 Experimental results

	Layers:5,1 Bins:6 Epch:2500 Pairs:2500 **(Exp-5.1a)**	L:5,5,51 B:6 E:2500 P:2500 **(Exp-5.1b)**	L:5,5,51 B:6 E:5000 P:5000 **(Exp-5.1c)**	L:5,5,51 B:31 E:5000 P:5000 **(Exp-5.1d)**
P@10	0.2909	0.3	0.3	0.2591
P@20	0.3182	0.3023	0.2932	0.2682
P@30	0.3152	0.303	0.3136	0.2879
NDCG@10	0.2174	0.2364	0.2446	0.1881
NDCG@20	0.2386	0.2339	0.237	0.1942
NDCG@30	0.2411	0.2386	0.2448	0.211
MAP	0.1461	0.1414	0.139	0.1331
Time	02:51	04:22	11:19	11:09

5.2 Parameters

drmm:numHistogramBins=6/31

drmm:mlpLayers=5,1/5,5,1

drmm:numEpochs=2500/5000

drmm:numTrainingPairs=2500/5000

drmm:learningRate=0.01

drmm:randomSeed=11642

5.3 Discussion

The deeper the layer does not result in better performance in our experiments. Since the network is more complicated now, it is possible it is not well trained. Therefore, in exp-c, I increase the training epoch and training data, while keeping the model the same (same bin and same layers). However, the performance does not improve. Increasing data and training time does not improve the performance of a more complicated model. Finally, I tried to increase the model complexity by increasing the bin to 31. Use more training epoch and data. However, the performance even decreases. My speculation is that the data is not diverse enough, so the model does not learn well even we train on more data and more time. This might only result in overfitting the model.

6 Experiment: Low vs. High Resource Systems (11-742 Only)

6.1 Experimental results

Provide information about the effectiveness of the "low resource" and "high resource" systems you developed.

Your .zip / .tgz file must include files named HW6-Exp-6.1a.qry, HW6-Exp-6.1a.param, etc., in the QryEval directory. The experimental results shown above must be reproducible by these files and the parameter values shown in the table.

	Low_1 (Exp-6.1a)	**Low_2 (Exp-6.1b)**	**$High_1$ (Exp-6.1c)**	**$High_2$ (Exp-6.1d)**
P@10	0.00000	0.00000	0.00000	0.00000
P@20				
P@30				
NDCG@10				
NDCG@20				
NDCG@30				
MAP				
Time	mm:ss	mm:ss	mm:ss	mm:ss

6.2 Parameters

Document the DRMM parameter settings that were used to obtain these results.

6.3 Discussion

Analyze the experimental results.