### 4.2 Evaluation of Purchase Probability Estimations

We first compare the purchase probability estimation accuracies between our models and the baseline. We use one day's log data for training and the next day's log data for test. Both the training and test data contain around 17 million records. And on average there are about 50 items in each record. So there are about 850 million items in both training and test data. For the baseline DNN model each item is a sample. For our models each record is used as a sequence of items, with each item being a sample. We use the Area Under the ROC Curve (AUC) and Relative Information Gain (RIG) [Chen and Yan, 2012] metrics for evaluation. The results on test data are shown in Table 1.

| Models | AUC | RIG |
|---|---|---|
| DNN | 0.724 | 0.094 |
| miDNN | 0.747 | 0.119 |
| miRNN | 0.765 | 0.141 |
| miRNN+attention | **0.774** | **0.156** |

Table 1: The test results of purchase probability estimation.

From Table 1 we see that the AUC and RIG of miDNN are much better than the baseline DNN. Note that the major difference between DNN and miDNN is that miDNN uses the global feature extension in Section 3.1 whereas DNN does not. This indicates that the global features are really useful and mutual influences between items are important indeed. By considering orders, miRNN has better results than miDNN. The best results are achieved by miRNN+attention, with a noticeable improvement over the results of miRNN. To see why miRNN+attention has better results, we visualize the attention $\alpha_{ij}$ from Equation (11) as follows. We find all test records whose length is at least 20, and calculate the $\alpha_{ij}, i \leq 20, j < i$. Then we average the values of $\alpha_{ij}$ from different records to obtain $\overline{\alpha}_{ij}$. So $\overline{\alpha}_{ij}$ is the average attention of position $i$ to position $j$. The $\overline{\alpha}_{ij}$ for $i \leq 20, j < i$ are shown in Figure 2, where each row corresponds to an $i$ and each column to a $j$.

From Figure 2 we see the top ranked items have larger attentions even when $i = 20$. This means that our RNN with attention learns to consider the influences from top ranked items even when estimating the probability of an item ranked far below. So the attention mechanism indeed alleviates the long-distance dependency problem that we stated in Section 3.2.

### 4.3 Online A/B Test

We also performed online A/B test to compare the GMV of our models to the baseline. In online A/B test, users, together with their queries, are randomly and evenly distributed to 30 buckets. Each experimental algorithm of ours is deployed in 1 bucket. And the baseline algorithm is deployed in the baseline bucket. The GMVs of experimental buckets and the baseline bucket are compared. Our online tests lasted for a time of a month to accumulate reliable results.

For each query, usually thousands of items need to be ranked. But statistics show that over 92% users browse no more than 100 items. And the average number of items browsed by users is 52. This means that only the top ranked items really matter. And as we stated in Section 3, the time complexity of miRNN and miRNN+attention are $\Theta(kN^2)$ and $\Theta(kN^3)$. For practical considerations on computing efficiency, the RNN models should not rank too many items. Therefore, we use our models in a reranking process in online tests. Specifically, we rerank the top-$N$ items of baseline ranking using our models. And we call $N$ the **rerank size** of our models. We experimented with different rerank sizes and compare the GMVs of our models to the baseline. The beam sizes of our RNN models are set to 5 unless stated otherwise. The relative GMV increase of our models over the baseline is shown in Figure 3.

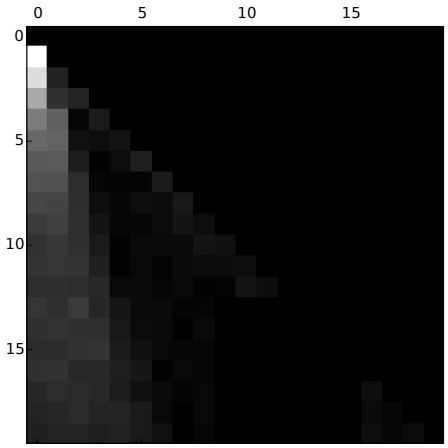


Figure 2: The average attention matrix when $N = 20$.

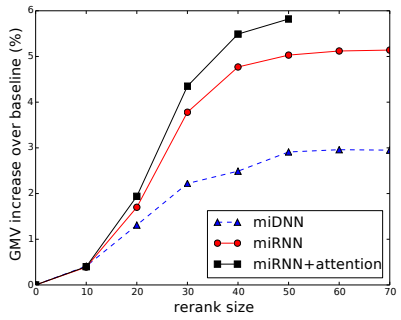


Figure 3: The GMV increase with respect to rerank size.

Figure 3 shows that our models increase GMV significantly over the baseline. The GMVs of our models increase as rerank size grows, but gradually stabilize as rerank size gets larger than 50. This may be explained by the statistics that 82% users browses no more than 50 items. So the benefit of increasing rerank size gradually gets small. Note that the maximum rerank size of miRNN+attention is limited to 50 for computing efficiency. To study the additional computational