

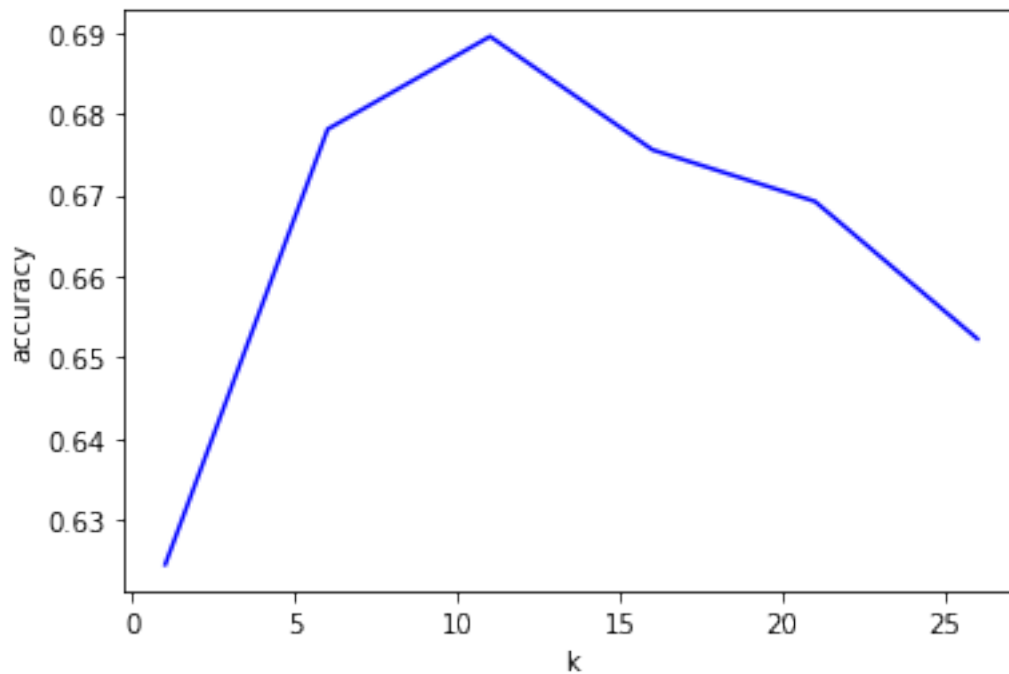
Final Project Q1

December 16, 2020

1 Q1

1.1 (a) (b) User-based knn

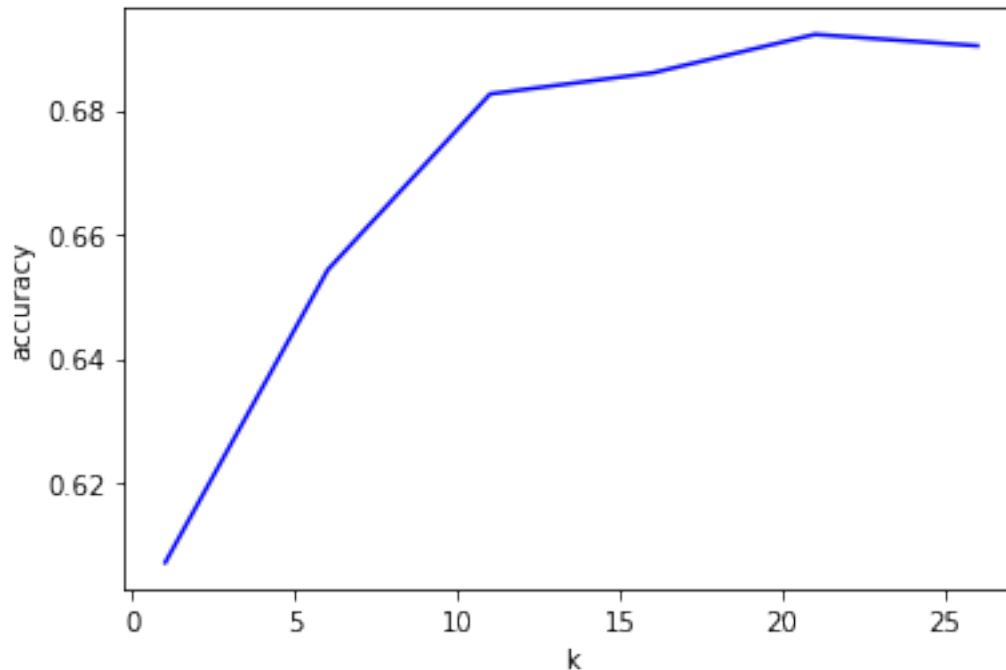
```
[5]: from part_a.knn import *  
import sys  
sys.path.append("../")  
  
sparse_matrix = load_train_sparse("./data").toarray()  
val_data = load_valid_csv("./data")  
test_data = load_public_test_csv("./data")  
  
k_list = [1, 6, 11, 16, 21, 26]  
user_based_report_knn(sparse_matrix, val_data, test_data, k_list)
```



The chosen k is 11 and the test accuracy is 0.6841659610499576

1.2 (c) Item-based knn

```
[4]: k_list = [1, 6, 11, 16, 21, 26]
     item_based_report_knn(sparse_matrix, val_data, test_data, k_list)
```



The chosen k is 21 and the test accuracy is 0.6816257408975445

1.3 (d) Comparison

1. The test accuracy from user-based collaborative filtering is slightly better than test accuracy from item-based collaborative filtering.
2. Besides, the computational time that user-based algorithm takes is much less than the computational time that item-based algorithm takes.

Therefore, user-based collaborative filtering performs better.

1.4 (e) Potential Limitations

1. The computation is expensive. The shorter one (user-based knn) still takes a long time compared to other algorithms (e.g. irt in q2).
2. The dimension of data is high, and in high dimensions, most points have approximately the same distance. So the nearest distance might not be useful.