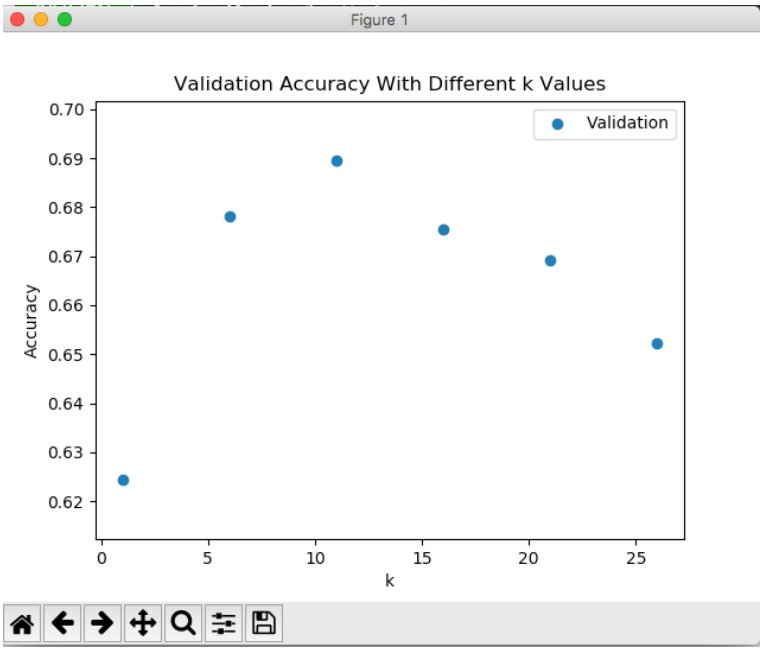
CSC311 Final Project

Nathan Chao

November 2020

1 Question 1

a)



k = 1 Validation Accuracy: 0.6244707874682472
k = 6 Validation Accuracy: 0.6780976573525261
k = 11 Validation Accuracy: 0.6895286480383855
k = 16 Validation Accuracy: 0.6755574372001129
k = 21 Validation Accuracy: 0.6692068868190799
k = 26 Validation Accuracy: 0.6522720858029918

b)
The k with the highest validation accuracy was $k = 11$ so choose $k^* = 11$. The final test accuracy with $k^* = 11$ is 0.6841659610499576.

c)
The core underlying assumption is that if question A has the same correct and incorrect answers by students as question B, A's correctness by specific students matches that of question B.

k = 16 Validation Accuracy: 0.6860005644933672
k = 21 Validation Accuracy: 0.6922099915325995
k = 26 Validation Accuracy: 0.69037538808919

The k with the highest validation accuracy was $k = 21$ so choose $k^* = 21$. The final test accuracy with $k^* = 21$ is 0.6816257408975445.

d) On test data, user-based collaborative filtering performs slightly better.

e)
One limitation of kNN is The Curse of Dimensionality. The data has many dimensions - 542 students and 1774 diagnostic questions. Because there are so many dimensions, nearly all data points will be "far away" from each other. A second limitation of kNN is its space complexity. We need to store the whole data set in memory. In comparison to many other machine learning algorithms in this course, this is bad space complexity. Another limitation of kNN is its time complexity. For every test sample, we must compare it to each sample in the training set. In comparison to many other machine learning algorithms in this course, this is bad time complexity.