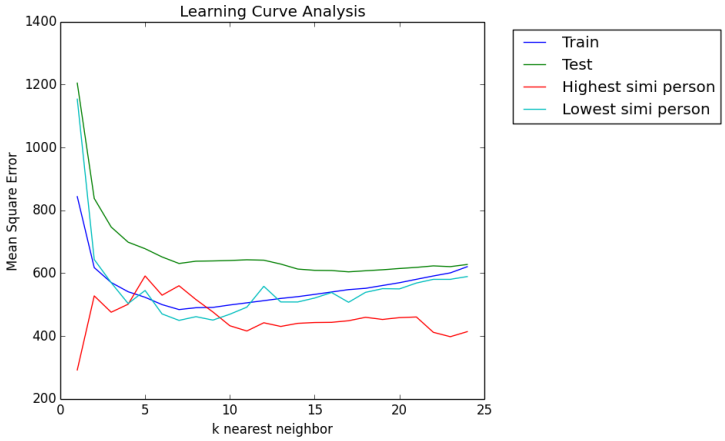
Info 4940: Algo we live by

ys849, yc2433

March 5, 2018

Assignment 1

Step 1: We first extract data and separate data into 25X19 data's training set and 25X1 data's testing set.
Step 2: We calculate similarity of each pair of person, find index of person that has highest (index = 21) and lowest(index = 2) mean similarity.
Step 3: We use knn_prediction function to calculate mean square error for training data, testing data, and also get mean square error for highest and lowest similarity person according to k(1-24) nearest neighbors. Step 4: Iteritavely run step 1 to step 4 19 times.
Step 5: Then, plot data we get as follow:



Observation:
First of all, we can see training mean square error decrease when number of nearest neighbor increase and at some point near 5 it increases afterward. For k = 1, MSE = 843.5. For k = 5, MSE = 523.17. For k = 24, MSE = 620.56. The reason of that is too large number of nearest neighbors will cause overfit, which makes mean square error increase.
As for test curve, its MSE always decrease when k increase. For k = 1, MSE = 1204.766. For k = 5, MSE = 677.76. For k = 24, MSE = 627.62. The reason for it might be number of training data is too small, so it cannot be accurately

predicted.
As for highest mean similarity person, we can see curve is roughly flat. For k = 1, MSE = 291.95. For k = 5, MSE = 591.066. For k = 24, MSE = 413.851. That means number of nearest neighbors doesn't influence predication too much. Moreover, highest mean similarity person seems to prefer less nearest neighbors to predict their rate.
As for lowest mean similarity person, we can see curve decrease sharply when k increase. For k = 1, MSE = 1153.75. For k = 5, MSE = 544.9. For k = 24, MSE = 589.1. That means prediction for dispersion in taste similarity person needs more neighbors.