Take home Quiz

1. [1.5 points]What is a possible disadvantage of using Pearson Correlation for finding missing ratings? Which extension to Memory-based algorithm can be applied to overcome this?
   Pearson Correlation considers only co-rated items. This may sometimes neglect the global behavior reflected in a user's entire rating history.
   Using Default voting, we can limit the weight such pairs contribute while computing the final prediction.
2. [1 points] How can you use clustering algorithms to produce a better result for recommendation systems?
   Clustering can help to identify similar sets of users/items in a more scalable way. The more expensive computations like pearson, etc can then be applied on this small subset instead of the entire range of similar users.
3. [1.5 point] What are the Characteristics and Challenges of Collaborative Filtering with Scalability? name two approaches to deal with this problem?
   Collaborative filtering systems are adversely affected by scalability. In cases such as Amazon, you can have millions of items and users, which could lead to lots of computations and time overhead at that big a scale.

   Approaches:
   Dimensionality reduction
   Clustering CF
4. [1 point] Give 2 cases with diagrams, where the BFR algorithm will fail, but the CURE algorithm will work.
   a. Cluster with tilted axis
   b. Clusters forming a shape similar to Yin-Yang
      (Both of these are there in the slides)
5. [3+2 points] Consider we have 2 clusters. Cluster A consists of the points P1(1,2,1), P2(2,1,3) and P3(0,3,2) and Cluster B consists of points Q1(4,5,7), Q2(6,6,6) and Q3(5,7,4). How will both of these clusters be represented if we are to apply the BFR algorithm?
   Now, consider the following points. We have to apply the BFR algorithm to determine what should be the assignment for each of these points. Computations of each point should be done independent of each other. We add a point to a cluster with a threshold of 3 standard deviations from the centroid. For this problem assume standard deviation is root(3) or 1.732

   1. Point X1(3,5,6)
   2. Point X2(3,5,4)

   Cluster A:
   N = 3
   SUM = (3, 6, 6)
   SUMSQ = (5, 14, 14)

Cluster B:
N = 3
SUM = (15, 18, 17)
SUMSQ = (77, 110, 101)

Point X1:
Distance from A: 6.57
Distance from B: 2.859
Hence X1 will be assigned to cluster B

Point X2:
Distance from A: 5.0371
Distance from B: 3.0394
Hence X2 will be assigned to cluster B
A is also within range, but B has a smaller distance.

Some common mistakes from students:
Not describing the clusters in the way mentioned above. This is clearly given in the PPTs

Using variance instead of standard deviation in the formula of Mahalonobis Distance.
This has led to larger distances and hence wrong answers.