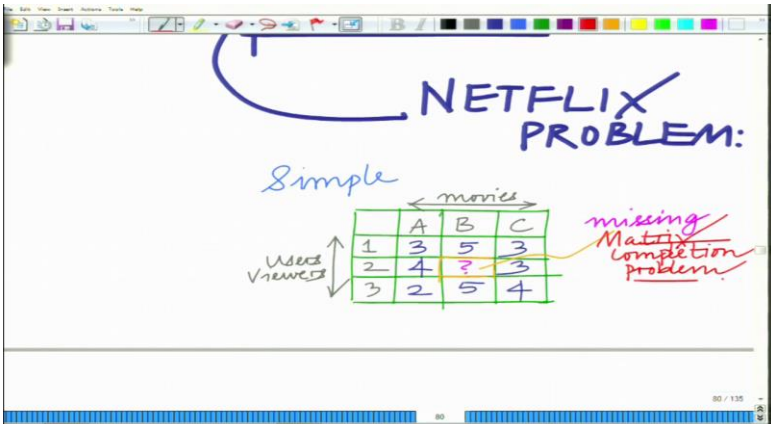
Applied Optimization for Wireless, Machine Learning, Big Data
Prof. Aditya K. Jagannatham
Department of Electrical Engineering
Indian Institute of Technology, Kanpur

Lecture - 78
Matrix Completion Problem in Big Data: Netflix-I

Keywords*: Matrix Completion Problem*

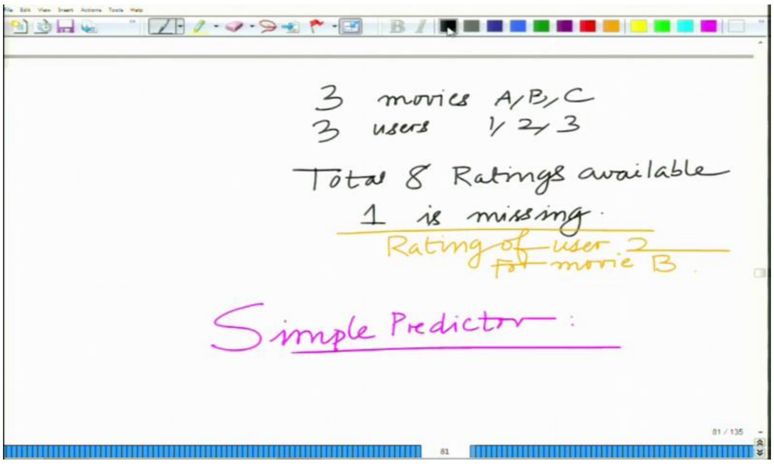
Hello, welcome to another module in this massive open online course. So we are looking at an application of Convex Optimization Big Data in particular to the Netflix problem. Let us consider a very simple version of that problem.

(Refer Slide Time: 00:48)



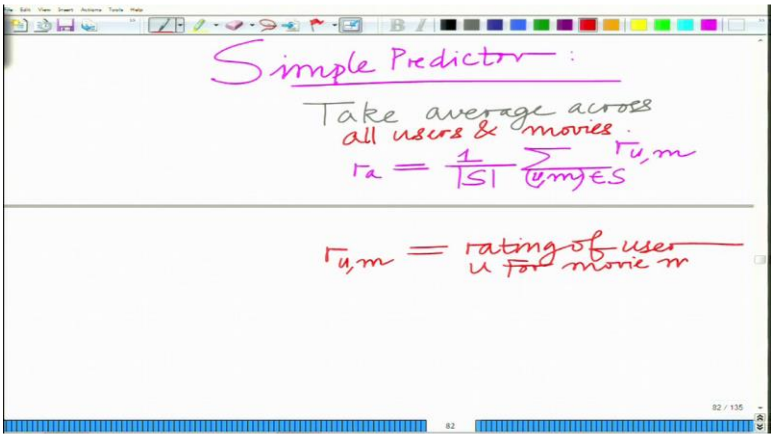
The Netflix problem as I said involves about half a million users and about 20000 movies. We are going to consider an extremely simple version of that which can be generalized. Let us say that you have 3 movies A B C and 3 users 1 2 3, similar to the table that we had seen yesterday. Let us consider again a simple example, where user 1 has rated movies A B C, the ratings are 3, 5, 3. User 3 has also rated movies A B C, ratings are 2, 5, 4. But user 2 has only rated movies A and C and his ratings are 4 and 3. Now, what you can clearly see is that one rating is missing, if you look at this as a matrix, this is missing. And therefore, we have to predict this rating to complete this matrix of users and ratings that is why it is also known as a matrix completion problem.

(Refer Slide Time: 03:24)



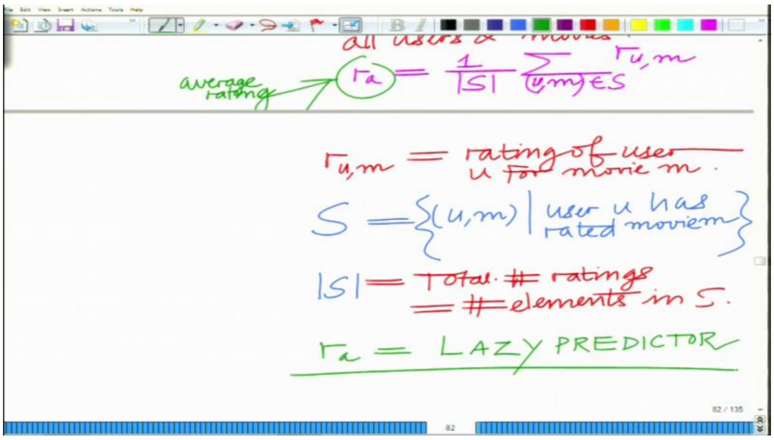
There are 3 movies A B C and 3 users. In total 8 ratings are available, one is missing, that is the rating of user 2 for movie B and this which we have to predict. Now let us consider a simple predictor and the best predictor is nothing but the mean that is the average rating of each movie per user.

(Refer Slide Time: 05:12)



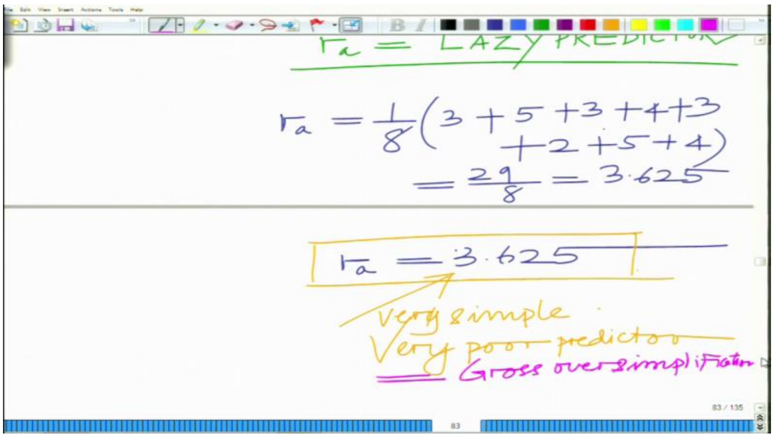
So we have rating $r_a = \frac{1}{|S|} \sum_{(u,m) \in S} r_{u,m}$ where S is the set of all ratings that are available. So $r_{u,m}$ is the rating of user u for movie m.

(Refer Slide Time: 06:28)



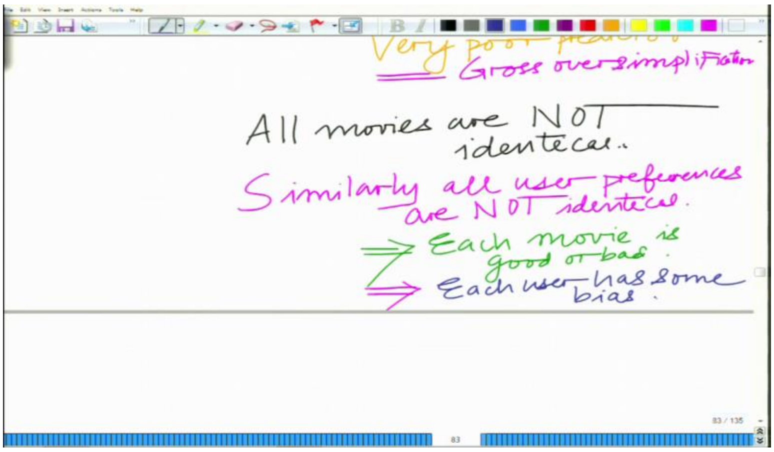
Now this is also termed as a Lazy Predictor as this is only the average rating.

(Refer Slide Time: 08:01)



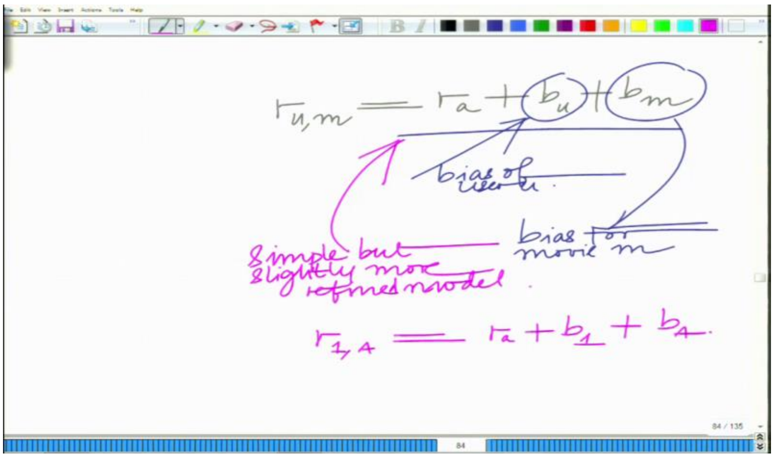
And for this example, the average is r_a = 3.625 as shown in slide. Its performance is going to be very poor because this is a gross oversimplification.

(Refer Slide Time: 09:17)



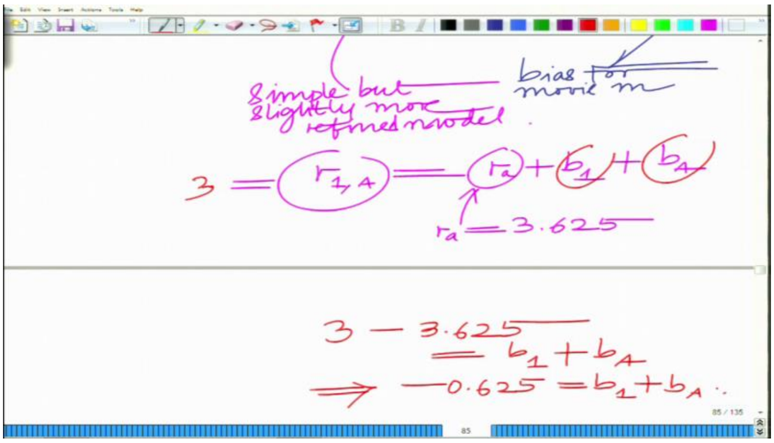
The reason for that is each movie is unique. So each user has a certain bias, each movie is good, but by and large, a large number of people think some movies are better than some other movies. So some movies are consistently rated the best. Some users are same way more lenient towards their ratings while some users might be very harsh. So to each user each movie is either good or bad.

(Refer Slide Time: 12:09)



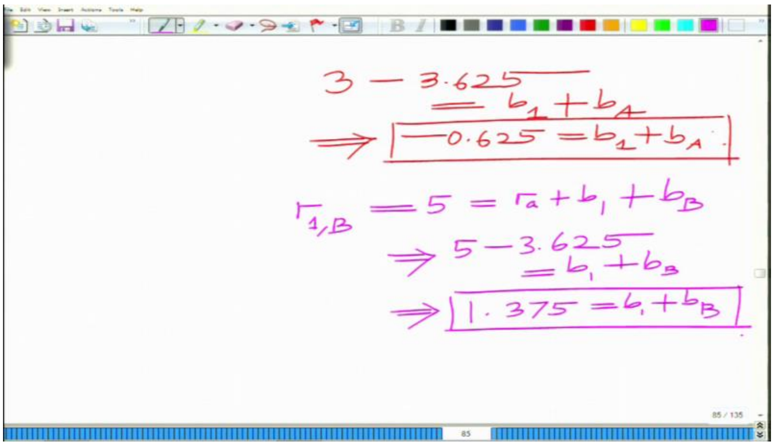
So we can model the rating of each user as $r_{u,m} = r_a + b_u + b_m$ which is the average, the bias of each user and the bias of each movie respectively. So this is a slightly more refined model and probably a more natural model in capturing the behaviour.

(Refer Slide Time: 13:46)



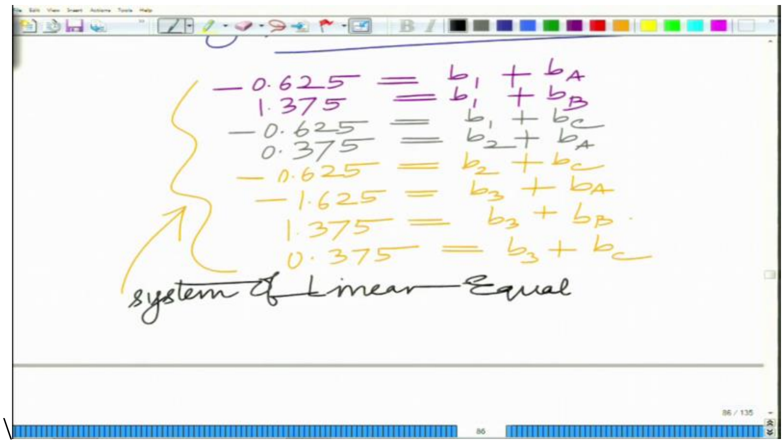
Accordingly, we can have various equations for the above problem as shown in slide.

(Refer Slide Time: 14:46)



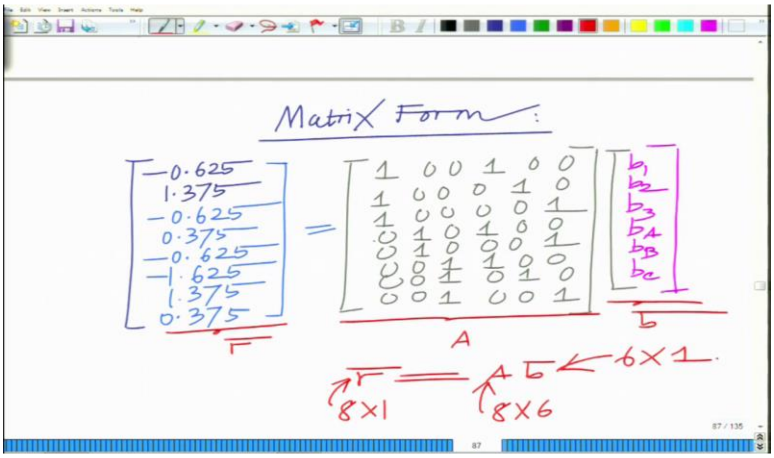
So there are 8 such equations as shown in slide.

(Refer Slide Time: 15:54)



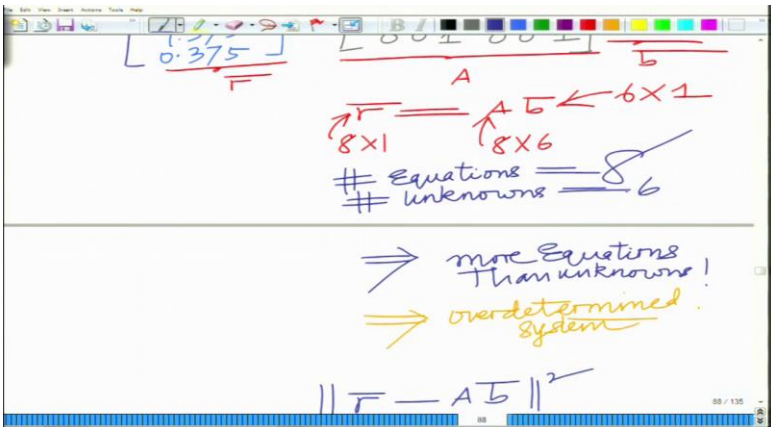
In fact, this is a system of linear equations which can be written in the form of a matrix as shown in slide.

(Refer Slide Time: 18:29)



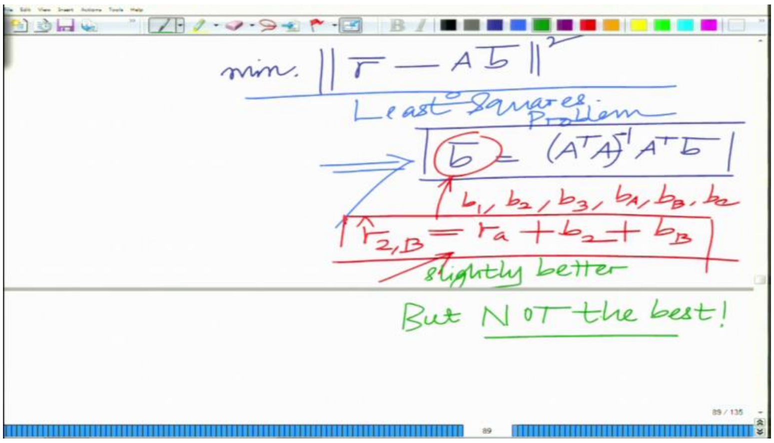
So here we have more equations than unknowns.

(Refer Slide Time: 20:38)



We know that this is an over determined system and hence we solve this over determined system by using least squares $\left\|\bar{r} - A\bar{b}\right\|^2$ but we cannot solve this exactly and we have to solve it approximately. So we find the best vector $\bar{b}$ which minimizes the norm of the approximation error $\left\|\bar{r} - A\bar{b}\right\|^2$.

(Refer Slide Time: 21:49)



So this is your least squares problem implies $\bar{b} = \left(A^T A\right)^{-1} A^T$. So from this you get b_1, b_2, b_3, b_A, b_B, b_C. So we have got first stage, we started with the average, we refined it, but we are not done yet. We are going to refine this model further to predict it as closely as possible and that we are going to do in the subsequent module. Thank you very much.