ISYE 8803 Exam 2 Problem 2

Nick DiNapoli, ndinapoli6@gatech.edu

July 31, 2022

1 Problem 1

In this problem, I analyze restaurant ratings of 300 different restaurants and run experiments to determine how missing data affects the performance of matrix completion. In this dataset restaurant ratings range from 1-5 and there are 200 measurements per restaurant. Specifically, I vary the amount of missing data from 5-50% using two different methods: missing data completely at random and missing data using a logical approach. For each method, I examine the reconstruction error as a function of missing data percentage and view the histogram of ratings in the original dataset, missing values dataset, and recovered dataset.

1.1 Missing values completely at random

First, I randomly select entries of the rating matrix to label as missing. I complete this by drawing random variates from a binomial distribution where the success rate (missing data rate) is set to a value ranging from 5-50%. The entries that get selected for missing data get set to zero. Figure 1 shows the result for the reconstruction error as the percentage of missing data is increased. As expected, I see that the error does in fact increase as the amount of missing data increases.

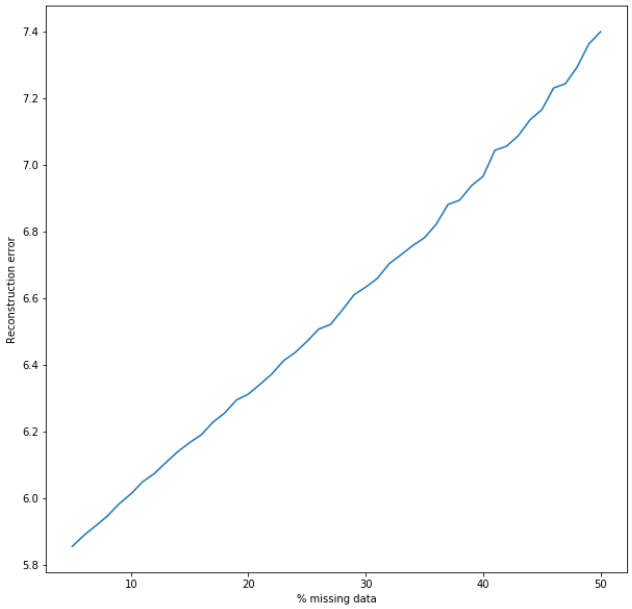

Figure 1: Reconstruction error for the first experiment.

Figure 2 shows the histograms of rating values for the matrices mentioned in the previous section when the amount of missing data is 50%.

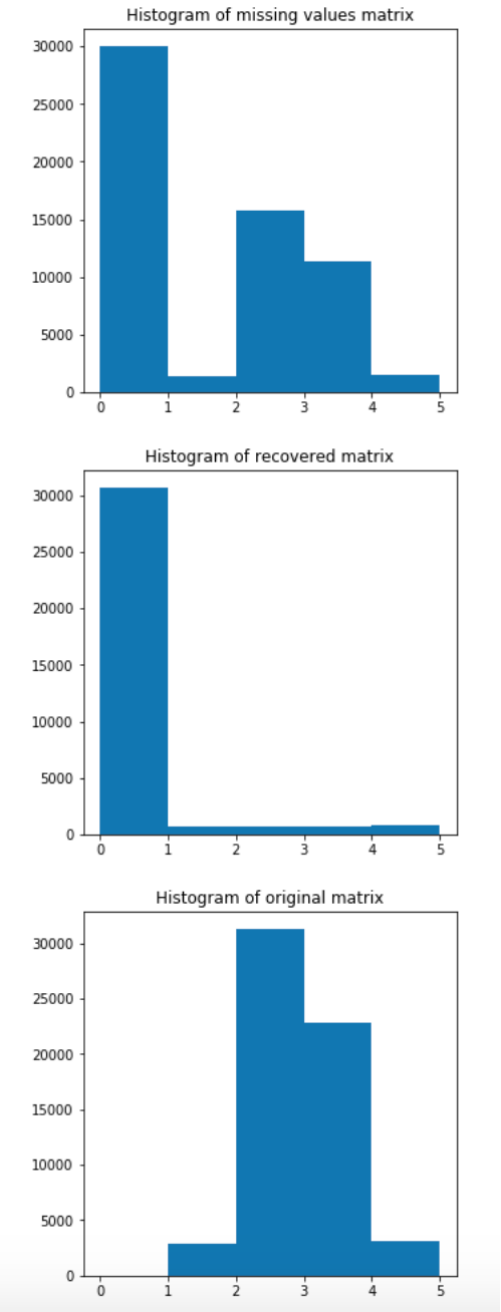

Figure 2: Histogram for the first experiment.

1.2 Missing values not at random

For the next experiment, I conduct the exact same task as the previous section except entries of the ratings matrix are dubbed missing in a logical manner. To simulate a customer avoiding a poorly rated restaurants, entries with a rating of a 1 or 2 are chosen to be three times more likely to be omitted compared to those with a rating of a 3 or 4. Entries with a 5 rating have a missing data rate of 0%. The success rates for the bionomial distribution are now adjusted such that the total missing data percentage is varied from 5-50%.

Figures 1 and 2 show the same results as the previous section but for this new experiment. It is clear that the more logical approach to missing data yields a lower reconstruction error.

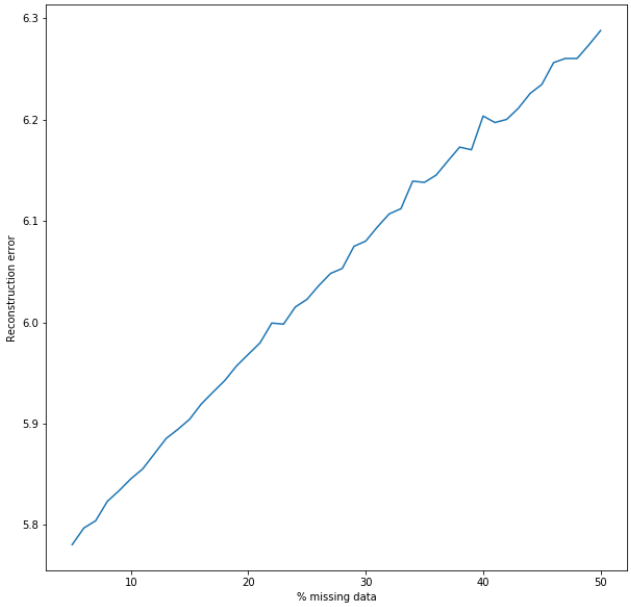

Figure 3: Reconstruction error for the second experiment.

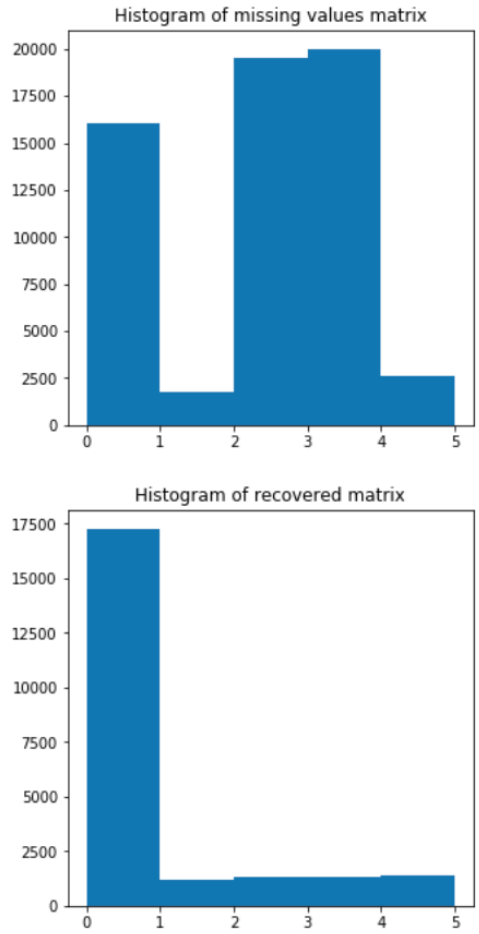

Figure 4: Histogram for the second experiment.