Experiment Design for Data Science - Exercise 1

Helmuth Breitenfellner, 08725866

9.11.2019

Data Exploration

The data contains user ratings for movies, together with data about the movies and data about the users.

Correlations: For the initial data exploration I have been using R. In a first step I denormalized the data into one data frame, added a few calculated fields (month, day of week or month, hour of review, movie age = time between review and release of the movie) and looked for correlations.

A few correlations could be identified. They are generally low, however given the number of samples some of them might be statistically significant (*I did not perform any specific statistical tests whether these correlations are statistically significant.*):

- Rating and movie age correlate with 0.17 (Pearson correlation coefficient)
- Movie age and reviewer age correlate with 0.12
- Hour of review and reviewer age correlate with 0.17

Also some small correlation between the month of the review and the rating (correlation coefficient = 0.04) was detected.

Then I looked into the histograms for the rating depending on the genre of the movie. Due to the page limit I did not depict them here, but some aspects are visible:

- Animation movies get the best, and Action movies the worst rating

Privacy Issues: The user related data is quite detailed. This could even be used to identify a single person. For example there is only one kid of age 7 in the user list - if it would be known that he is in the list one could clearly identify his ratings.

Hypothesis A

The correlation between rating and month of the rating can be used to improve the performance of a User-User CF rating prediction algorithm.

Dependent and independent Variables The *independent variable* in this hypothesis is whether an adjustment of the rating depending on the month of the review shall be appied or not.
The *dependent variable* is the performance of the prediction algorithm.

Control Condition In the first case, a User-User CF rating prediction is performed and the performance is measured.

In the second case, the influence of the month of review for the known ratings is first substracted. Then from the adjusted input the User-User CF rating prediction is performed. At the end the influence of the month of the review is added to the predicted rating.

The User-User CF rating prediction algorithm is in both cases exactly the same. All other aspects (e.g. data set sampled) are the same as well.

Performance Indicator For measuring the performance RMSE is used. A statistical test (e.g. *sign test*) is used to decide whether the number of cases with increased performance is statistically significant.

Simulate Real-World Conditions The test data is later in time than the training or development data.

Hypothesis B

Using half-precision (16 bit) floats in an SVD algorithm will not affect the performance of the rating prediction when compared with double precision (64 bit) floats, but will significantly reduce the runtime.

Dependent and independent Variables: The *independent variable* in this hypothesis is whether half-precision (16 bit) or single precision (32 bit) calculations are performed.
The *dependent variables* are the prediction performance as well as the runtime of the prediction algorithm.

Control Condition: In the first case, the SVD algorithm is used to predict ratings from previous ratings. For this the algorithm is running using 32 bit floats. The implementation is written in CUDA. The experiment is performed on a Tesla V100 GPGPU.

In the second case, the same SVD algorithm is used; however, the numerical precision is changed to half-precision (16 bit).

The SVD rating prediction algorithm is in both cases exactly the same. All other aspects (e.g. data set sampled) are the same as well.

Performance Indicator: For measuring the prediction performance RMSE is used. A statistical test is used to decide whether the difference in prediction performance is significant or not.
The runtime is measured as elapsed total time, including time for file I/O. A statistical test is used to decide whether the difference in runtime performance is significant or not.

Simulate Real-World Conditions: The test data is later in time than the training or development data.

Hypothesis C

A gender-sensitive rating prediction is performing better in ranking movies.

Dependent and independent Variables: The *independent variable* in this hypothesis is whether only gender conforming data is used or all data irrespective of gender.
The *dependent variables* is the prediction performance.

Control Condition: In the first case, a state-of-the-art rating prediction algorithm (e.g. again the SVD algorithm) is used to predict a rating for a user-movie combination, using all available data in the training set.
In the second case, the same algorithm is trained twice: once for ratings from men and once for ratings from women. This means that the effective training set is smaller for the two models, but the models are better aligned to the target used. All other parameters are kept the same. To predict a new rating, the model is chosen depending on the gender of the user whose rating is to be predicted.

Performance Indicator: For measuring the performance RMSE is used. A statistical test (e.g. *sign test*) is used to decide whether the number of cases with increased performance is statistically significant.

Simulate Real-World Conditions The test data is later in time than the training or development data.