

Project Documentation

Mat-Nr.: 6574933

Birthday: November

Motivation and Introduction


























My Birthday is in November so my focus is "Data Analysis - Explainability - Performance", so I choose the Recommender System for Movies-Task because I think Data Analysis works better for this task. In addition, I find the task of a recommender system much more interesting, since such systems meet us everywhere in daily life: Whether with Netflix or YouTube or with advertising we receive.

I used the given cold start as the starting point. However, it uses a reduced data record without demographic user information. Accordingly, a model is used which is only trained with the UserID.

My intention was to expand this model with demographic data such as age, gender or occupation. For this purpose, the complete MovieLens data record had to be read from csv-Files and reshaped so that it is possible to train the model with this data. This step led to major challenges because there were errors with data types and lengths of user and movie elements.

Modell

The model used uses collaborative filtering, which is a widely used technique for recommender systems. It uses the behavior and evaluations of other users to predict the evaluation of a user. In doing so, users are searched for who have similar interests and their evaluation of the required size. This concept can be seen in the figure on the right. The users who interact similarly are green highlighted. The prediction will be negative because the similar users have given a negative evaluation of the size searched for.

Collaborative Filtering

https://en.wikipedia.org/wiki/File:Collaborative_filtering.gif

Konzept

The model from the example, which only receives UserID and MovieID, serves as a base or reference model. For the extended model, the given was extended by the user information age and gender. Besides, genre information from the movies got also included. The full model also receives the occupation of the user and the timestamp of the evaluation.

Another comparison model is one in which the UserID has been removed. It only learns with the demographic data of the user. Thus, the influence of this data on the result is to be tapped.

In the first step, the pre-built training and test sets of MovieLens were used as data sets. Subsequently, a separate data split was also performed and compared. Cross validation was also carried out and its influence on the result was considered.

Analyse

The Base, Extended, and Full models all deliver the same validation loss with the loss function "BinaryCrossentropy". With the loss function "MeanSquaredError", there are some differences in the results. Even the movie recommendations differ only slightly, in some cases only in the order, in some cases there's no difference at all. Overall, they all provide realistic and similar results. The models also provide the same results when predicting films that have already been reviewed. Neither model is significantly better or worse than the other.

When comparing the extended model and the cross validation model, it is noticeable that the recommendations are rather different. Above all, the sequence deviates more strongly. The differences in prediction for viewed films are also bigger. The values from the cross validation tend to differ more from the real value. Therefore, the larger validation loss also seems sensible.

There seems to be overfitting when training with my own data split. The validation loss is significantly greater than that of the models trained with the pre-built sets. Further adjustment (then early stopping and reduce learning rate) of the training parameters in order to possibly prevent this overfitting did not take place in this project.

The model that was trained without UserIDs also has a larger validation loss than the base model only trained with UserIDs.

Outcome and Outlook

Same performance with advanced data was unexpected. My assumption is that the UserID behind a human/individual already contains more information than can be generated with demographic data. This also shows the worse result of the "Without UserID" model. The additional data, at least for this network, does not seem to provide any new information that contributes to performance improvement.

Different usecase: User Activity?

With another network, which is based on the model from the DeepDive tutorial, the user activity is to be predicted, measured based on the reviews given. Age, gender and occupation are to be used for this purpose. This should predict how active new users will be.

However, the problem for this task is that due to the mass of data and differences between users with similar properties, the values always move quite close to the mean value. For marginal groups such as the elderly, there is a significantly lower activity, which should also correspond to reality.