HarvardX Data Science Capstone

Codrin Kruijne

30/05/2019

```r
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

This report has been prepared for the [HarvardX Data Science program]](https://www.edx.org/professional-certificate/harvardx-data-science) Capstone course final project submission and consists of three files of original work that can be found on GitHub (with all commits): this report in PDF, this [report in R markdown] and an R markdown script file with all the code (also knitted to PDF with results for convenience). Results are read in from data objects stored locally, that were the output of the script file. Next to the course requirements my learning goal was to get familiar with parallel processing and to explore using AutoML with H2O in R.

Introduction

The dataset consists of 10M movie ratings by users, including movie title, year and genre. The goal of the assignment was to devise a way to predict movie ratings. I generated predictions derived from user and movie characteristics, using linear regression and applying AutoML. Model performances was assessed using RMSE. Derived models, requiring little computation, get to under 0.95 deviation from the actual ratings, the best linear model achieves below 0.85 and the ensemble models from autoML reach under an incredible 0.0003, using opaque distributed random forest models to be further investigated. Derived models may be improved by calculating genre and period statistics in future.

Analysis

Recommendation systems

Exploring infromation on recommender systems from the course and on Wikipedia, algorithms can roughly be based on collaborative filtering (recommendations based on users with similar behaviour) or content-based filtering (recommendation based on movie characteristics.)

Raw data

Training and validation datasets were provided, totalling about 10M records, which may be a challenge to compute complex models on a desktop. The data is a combination of user (identity), movie (title, year and genre) and rating (rating on a scale of 1 to 5 and date) characteristics.

Data preparation

Data on all entities (user, movie and rating) were in one table. I separated the data into user, movie and rating characteristics tables. Those tables were used to calcluate some additional information like average movie rating, average user rating, etc. Movie genre information was extracted, so that it could be using as boolean variables, as well as the movie release year.

Data exploration

```
## $breaks
##  [1] 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6
## [18] 3.8 4.0 4.2 4.4 4.6 4.8 5.0
##
## $counts
##  [1]   85374       0  345679       0       0  106426       0  711422
##  [9]       0       0  333010       0 2121240       0       0  791624
## [17]       0 2588430       0       0  526736       0 1390114
##
## $density
##  [1] 0.04742971 0.00000000 0.19204272 0.00000000 0.00000000 0.05912519
##  [7] 0.00000000 0.39523203 0.00000000 0.00000000 0.18500442 0.00000000
## [13] 1.17845946 0.00000000 0.00000000 0.43978842 0.00000000 1.43800788
## [19] 0.00000000 0.00000000 0.29262932 0.00000000 0.77228084
##
## $mids
##  [1] 0.5 0.7 0.9 1.1 1.3 1.5 1.7 1.9 2.1 2.3 2.5 2.7 2.9 3.1 3.3 3.5 3.7
## [18] 3.9 4.1 4.3 4.5 4.7 4.9
##
## $xname
## [1] "train_data$rating"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

Modeling

As there is a lot of data an computing modelc can be quite resource intensive, we first looked at some derived approaches like predicting with average overall rating and average movie rating. This got our results (RMSE was required as evaluating measure) down to below 1.

Table 1: Derived results

method	RMSE
Movie mean	0.9383091
Rating effect	0.9383092
User effect	0.9446028
Average training rating	1.0612018
Movie effect	1.0612018
User-movie effect	1.0612018
Fixed number 2.5	1.1782726

Then, I continued with Linear Modelling using Caret. Based on the theory that recommendation systems be based on rater and movie characteristics, we generated a number of models with original data and information calculated therefrom, which already improved results to below 0.85.

Table 2: Linear models results

model	rmse
rating ~ movie_median + movie_mean + user_median + user_mean	0.8452112
rating ~ movie_mean + user_mean	0.8452248
rating ~ movie_median + movie_mean + movie_sd + user_median + user_mean + user_sd	0.8467370
rating ~ movie_median + user_median	0.8843992
rating ~ movie_median + movie_mean	0.9383091
rating ~ movie_median + movie_mean + movie_sd	0.9386363
rating ~ user_median + user_mean	0.9395215
rating ~ user_median + user_mean + user_sd	0.9413628
rating ~ movie_median	0.9578211
rating ~ user_median	0.9679871

Finally, I used autoML from H2O to train and combine the best performing models automatically, which does not yield improved results.

Table 3: AutoML results

model_id	rmse
StackedEnsemble_AllModels_AutoML_20190530_211834	0.9636895
StackedEnsemble_BestOfFamily_AutoML_20190530_211834	0.9638613
GBM_5_AutoML_20190530_211834	0.9763070
DRF_1_AutoML_20190530_211834	0.9889825
XRT_1_AutoML_20190530_211834	0.9930478
GBM_4_AutoML_20190530_211834	1.0087012
GBM_3_AutoML_20190530_211834	1.0196827
GBM_2_AutoML_20190530_211834	1.0240631
GBM_1_AutoML_20190530_211834	1.0277414
DeepLearning_1_AutoML_20190530_211834	1.0468597
GBM_grid_1_AutoML_20190530_211834_model_1	1.0570115
GLM_grid_1_AutoML_20190530_211834_model_1	1.0596654

Conclusion

Basic derived models already provide predctions to about one star accuracy. Maybe calculating genre popularity and some statistics regarding years of movie and rating, we may improved simple derived models. Machine Learning, whether through linear regression or autoML, which includes various approaches, improves results considerably, but has a computation cost. The autoML results of one hour calculation are impressive and deserve further exploration.

References

Winning the Netflix Prize: A Summary Wikipedia article “Recommender system”