Machine learning project – Machine learning D CS-C3240

Book Recommendations with Linear Regression in Excel

1 Introduction

As an active book reader, it is often helpful to get book recommendations based on your reading preferences to get a better reading experience. In order to get a recommendation that matches your own preferences, it is important to have an understanding of what genres or themes you enjoy in books. Even smaller factors, such as book length and whether a book uses a more simple or complex language can affect if you enjoy a book or not. The aim of this machine learning project is to give possible ratings for books a person hasn't read, by looking at ratings given by the same person to previously read books, in order to define how much the person would enjoy a specific book. Section 2 will discuss how the problem was formulated and how data was obtained, section 3 will discuss the method used in detail and section 4 will discuss an evaluate the obtained results.

2 Problem formulation

Book recommendations can be modelled as machine learning (ML) problems with data points representing individual books. Each data point or book is characterized by different features such as author, genre or whether the book is a part of a series or a standalone novel. In this ML project the features will only be defined by the genres the book belongs to. The label or quantity of interest for a datapoint is how well a certain book is recommended for you based on your preferences. Datapoints were gathered from Goodreads.com [1], where a person can rate books they've read from a scale of 1-5.

In Goodreads each book is classified into a genre based on what users classify it as when reviewing a book. As a consequence, one book can be classified into many genres by varying "votes" instead of being classified into only one genre. This project used 9 different and common genres as features (Fantasy, Science fiction, Young adult, Adult, Romance, Fiction, Contemporary, Mystery and Classic) to describe each data point and only looked at the genre classifications with most votes for a book. For each single datapoint, the amount of user votes was divided with the genre that had most votes for that specific book. For example, if a book had 27 049 votes for being a science fiction novel and most votes for being a young adult book (31 318) this would correspond in the book being 0,863688

parts a science fiction book. If the book didn't have any votes into one of the 9 genres, the value for that feature was chosen zero.

The labels were then the personal ratings a user (me) had given for the books on a scale of 1-5. The first dataset with known labels thus consisted of 29 datapoints, with each row describing one datapoint or book. This dataset, called Dataset 1, was used to train and validate the three models used in this project. A second dataset, called Dataset 2, was collected from Goodreads.com with 9 datapoints with known labels and was used as a test set to check the quality of the two models. A third dataset with 15 datapoints, Dataset 3 with unknown labels, was used to see how the best model would function in order to recommend books.

3 Method

In order to find out whether a book with a personal high rating (label y) had any linear correlation with the genres (feature x) it belonged to, all of the models used in the project were linear predictors performed in Excel by linear regression. Dataset 1 was first divided into a training and validation set by a single random split with the Excel "RAND"-function. This resulted in 19 books being in the training set and 10 in the validation set. Weight vectors w_1-w_9 were chosen for each feature based on how much I enjoy a certain feature and the bias term w_0 was put to zero. The predicted rating h(x) for each book was calculated with the sum product for each feature and its corresponding weight while also adding the w_0 value into the final sum.

The quality of the hypothesis h(x) was then evaluated by calculating the training and validation errors for each datapoint. The loss functions were calculated with the squared error loss $(y-h(x))^2$, since squared error loss works well for ML problems involving numeric labels [2], by comparing the rating I had given with the hypothesis. After this the average training error and validation error were calculated, since linear regression learns a predictor that minimizes the average squared error loss [2]. In order to minimize the training and validation error, the weights had to be optimized. This was done with the Excel "Solver"-tool that minimized the training and validation error by optimizing the weights [3]. In this project the weights could be positive or negative.

Three models were tried out with the Excel "Solver"; Model 1 that regarded all of the 9 features and two that further optimized the hypothesis by only regarding some of the features. For Model 2 the

five first features were used, so only w_0 and w_1-w_5 were taken into account when optimizing the hypothesis. For the last model, named Model 3, only four features were looked at, thus optimizing only w_0 and w_1-w_4. The average training and validation errors for the three models used on Dataset 1 are presented in Table 1.

Table 1. Average loss for Dataset 1 with three linear regression models.

	Training error	Validation error
Model 1	0,520373	2,525115
Model 2	0,918826	1,017309
Model 3	0,925465	0,934066

4 Results

Based on the results in Table 3, Model 3 with the least number of features resulted in the smallest average validation error (0,934066). Model 1 is most likely overfitting based on how large the difference between the average training (0,520373) and validation error (2,525115) is. Moreover, the validation error for Model 1 is quite large. In contrast, Model 2 has a smaller average validation error (1,017309) and could be considered an okay model since it is quite similar to the training error (0,918826). Model 3 is the best one of the three models, since it has the smallest validation error and a small difference between validation and training error. It seems that the more features, the more the machine learning method overfits on the training data.

The results from the best model, Model 3, were applied to the test set, Dataset 2. Test errors for each datapoint were calculated with the squared error loss and the average test error was thus 0,88701809. Finally, I tried the best model for books with unknown labels from Dataset 3, in order to get a prediction on how much I would enjoy a book I haven't read. The predicted ratings for 4 out of the 15 datapoints are presented in Table 2. Ratings have been rounded to match the rating system of the training and validation set.

Table 2. Predicted ratings for books I have not read yet.

Book	Author	Hypothetical rating based on Model 3

Skyward	Brandon sanderson	4,0
Illuminae	Amie Kaufman & Jay Kristoff	3,0
Ballad of songbirds and snakes	Suzanne Collins	3,0
Elenor & Park	Rainbow Rowell	2,0

5 Conclusions

In this ML project linear regression in Excel was applied to create a system that could recommend books to me based on what genres they belong to. The three models used in this project were obtained by linear regression with different amounts of features taken into account. Squared error loss was applied on each model and the model with least features (Model 3) resulted in the smallest average validation error 0,934066 and average training error 0,925465. Overall, the results for Model 2 and Model 3 are good and do not seem to overfit, whereas Model 1 overfits. The best model was then tested on the test set (Dataset 2) that gave an average test error of 0,88701809. Model 3 was also tested on a dataset with unknown labels, Dataset 3.

The results obtained for Dataset 3 are somewhat aligned with my personal preferences, predicting that I won't enjoy contemporary fiction novels such as Elenor & Park by Rainbow Rowell as much as science fiction novels like Skyward by Brandon Sanderson. The results indicate that there is a linear relation between the books I enjoy and into which genres they belong into. Although the results of this ML project seem decent, it is worth noting that many important features were left out from the project, such as author, book length, year of publication and whether the novel is a standalone or series. As seen in the project, using many features lead to overfitting. This could possibly be improved with using more datapoints, since this project only used 29 for the validation and training set. Another approach would be to use a more complex model instead of linear regression, such as a recommender system.

References

1. Goodreads.com https://www.goodreads.com/book/show/2767052-the-hunger-games [Accessed 19th March 2021]

2. A., Jung. (2021). *Machine Learning: The Basics.* mlbook.cs.aalto.fi
3. EngineerExcel. Linear regression in Excel: 3 alternative methods https://engineerexcel.com/linear-regression-in-excel-3-alternative-methods/ [Accessed 22nd Match 2021]