Netflix movie recommendation system using alternate least square matrix factorization

Márcio Alexandre Silva Monteiro, Prof. Dr. Paulo Henrique Ferreira da Silva(Advisor)
Federal University of Bahia (UFBA)

Abbreviated abstract: The present work aims to study and create a movie recommendation system from a database made available by Netflix. For this purpose, the technique of matrix factorization by alternating least squares was used, due to its computational efficiency and availability in the MLlib library of Apache Spark. The performance of the proposed model was evaluated according to an appropriate error measure and an empirical evaluation was also carried out based on the result found.

Related publications:

– Izbicki e Santos, Aprendizado de máquina: uma abordagem estatística (2020)
– Ghosh *et al.*, Recommendation System for E-commerce Using Alternating Least Squares (ALS) on Apache Spark (2021)

marcioasm84@gmail.com

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022

Problem, Data, Previous Works

Why recommend?
Why create a recommendation system?

Data obtained from Kaggle:
https://www.kaggle.com/code/jieyima/netflix-recommendation-collaborative-filtering/data

- Features:
  - ✓ 100,480,507 scores (17,770 movies and 480,189 users)
  - ✓ Film Release Year (between 1896 and 2005)
  - ✓ Sparsity of the user-item matrix (93.65% zeros)
  - ✓ The highest number of scores given by a user was 17,436 and the lowest was 11
  - ✓ 1 Movie title archive (Movie_Id, Year, MovieName)
  - ✓ 17,770 movie scores files (Movie_Id, Cust_Id, Rating, Date)

- Netflix Prize Challenge (from 2006 to 2009)

marcioasm84@gmail.com

Methods

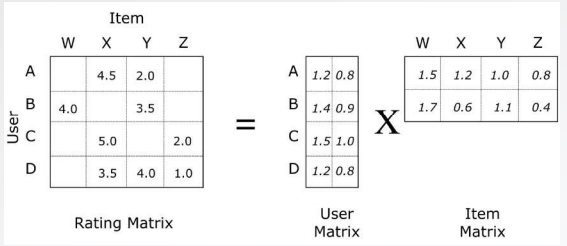
- Treatment of files to allow them to be read
- Big data
- Using Apache Spark to enable reading and machine learning MLlib

Alternating Least Squares (ALS):

- Model-based collaborative filter
- Transformation of the user-item matrix into the product of two other matrices of smaller dimensions
- Solution for: popularity bias, item cold-start problem, scalability issue

Hyper-parameter: maxIter =10, regParam =0.01, rank =10

Root-mean-square error (RMSE): 0.8516



Source:https://towardsdatascience.com/

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2},$$

yi : is the predicted value for the i-th observation;
xi : is the observed (actual) value for the i-th observation;
n : is the total number of observations;

marcioasm84@gmail.com

4th Conference on
Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Results and Conclusions

Low error value (RMSE = 0.8516)
Predicted Great Scores

Movie_Id	MovieName	Rating	Prediction
8743	Ice Age	5	4.5960402
976	Tom and Jerry: Paws for a Holiday	5	4.8389072
6001	Tom and Jerry: The Movie	5	4.8184667
16660	The Lion King II: Simbas Pride	5	5.1620736
3079	The Lion King: Special Edition	5	4.993923
16222	The Lion King 1 1/2	5	5.122994

Creation of user who only watched cartoons

Best recommendations:

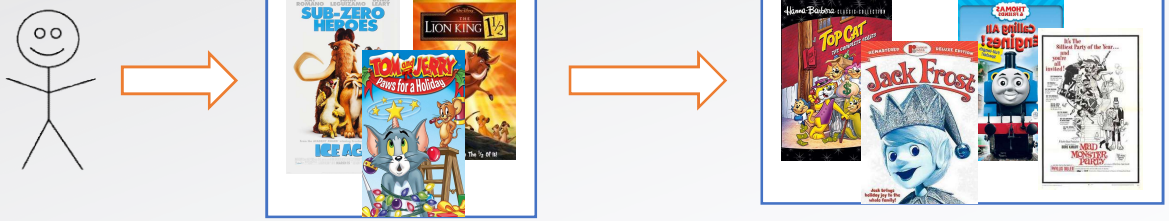
Top Cat: The Complete Series (Cartoon), Jack Frost (Animated) (Cartoon/Comedy),
Thomas & Friends: Calling All Engines (Cartoon), Mad Monster Party (kids/musical) ,
Goosebumps: The Ghost Next Door (family/fantasy/horror/mystery/thriller).

Other recommendations:

The River (Drama/Romance), National Geographic: Predators at War (Documentary),
Gospel (Music/Documentary), Jekyll & Hyde: The Musical (Musical/Horror),
Bob Hope: Hollywood's Brightest Star (Documentary),
Dark Shadows: Vol. 14 (Fantasy/Horror), The Fighting Sullivans (War/Drama)

Movie_Id	MovieName	Year	Rating
1605	The River	1997	7.2448993
14589	Top Cat: The Complete Series	1961	7.1064553
10807	Petticoat Junction: Ultimate Collection	1963	6.951481
8267	Do You Remember Dolly Bell?	1981	6.9319854
4941	Aetbaar	2004	6.807052
1868	Mad Monster Party	1967	6.798548
16072	Basic Ab Workout for Dummies	2002	6.693246
6106	The Jazz Channel: Chaka Khan	2000	6.685393
4665	Goosebumps: The Ghost Next Door	1998	6.67755

Total recommendations: 20



marcioasm84@gmail.com

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022