The Rec'ing Crew

Check Yourself Before You Rec Yourself©

Mickey Mulder

Paul Drabinski

Alex Worth

INTRO

Is it worth investigating other machine learning methods in order to optimize the Movie-Legit recommendation system?

Considerations?

- Accuracy
- Speed

Methods?

- Mean of Means
- Alternating Least Squares
- SVD and SVD++

Tests & Results?

Mean of Means

How it works

- Averages the average of the entire matrix, the average of each users ratings, and the average of each items ratings

RMSE (100,000, 1 million, 10 million) = 1.01, 0.997, 0.951

Pros

- It’s a simple method, easy to understand and troubleshoot

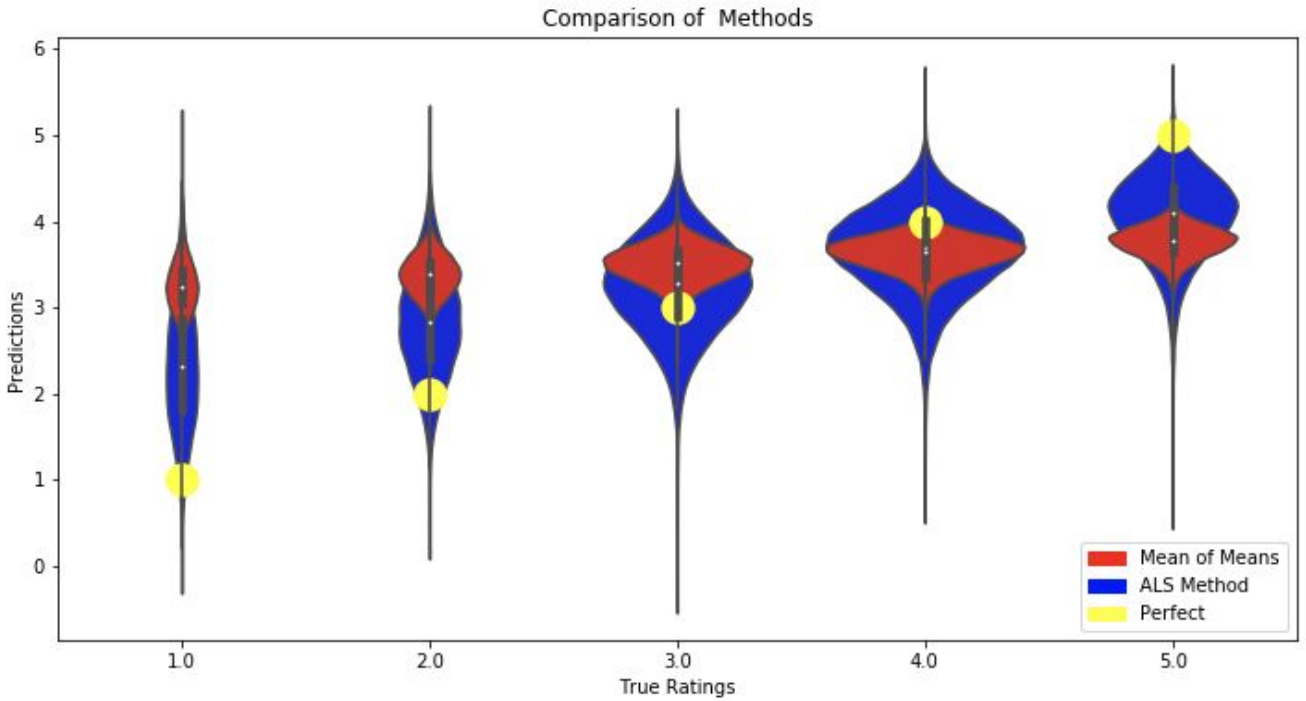
Cons

- Could be very inaccurate for certain users depending on how they rate (ie. User A only rates things at 5, but could get auto filled ratings at 3)
- Paleolithic era method

Alternating Least Squares

- Why it’s better?
  - Better Recommendations (1 million ratings dataset)
    - Provides a RMSE of 0.851.
    - This is a 14 .6 % increase in prediction accuracy over mean of means.
    - What does that mean? A prediction is 14.6% closer to the actual rating a user would have given it!

Alternating Least Squares



Alternating Least Squares

- Why it’s better?
  - Performance benefits with scaling (likely would need more testing)
    - Utilizes Apache Spark
      - Allows you to run programs on multiple machines
      - Easily parallelizable
      - The algorithm builds significantly better models with more data

Alternating Least Squares

	RMSE		Time to Run (s)	
	Spark_ALS	MOM	Spark_ALS	MOM
100,000	0.915	1.02	8.74	4.63
1,000,000	0.851	0.997	22.89	71.08

- All times compared on one machine, MacBookPro, 2.7 GHz Intel Core i7, 16 GB RAM

Alternating Least Squares

How it Works:

The *ALS* algorithm first randomly fills the users matrix with values and then optimizes the value of the movies such that the error is minimized. Then, it holds the movies matrix constant and optimizes the value of the user's matrix.

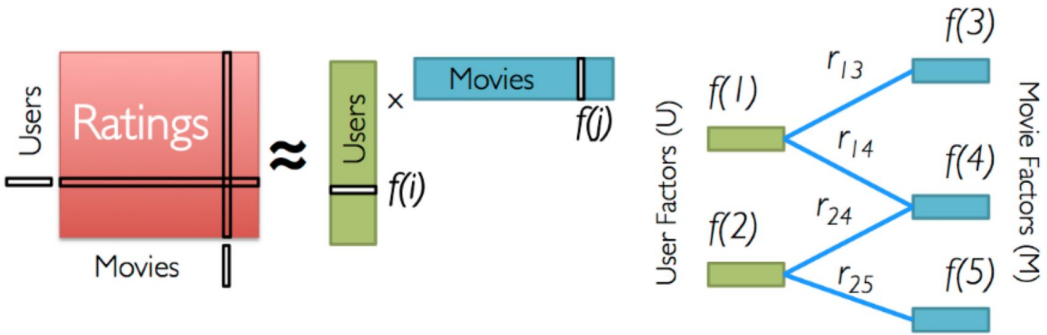


Image from : http://ampcamp.berkeley.edu/big-data-mini-course/movie-recommendation-with-mllib.html

Alternating Least Squares

- Drawbacks
  - Complicated- More parameters
  - Difficult to troubleshoot
  - There are potentially better machine learning algorithms (SVD++, SVD), that are more computationally intensive (either not parallelizable or not available on spark)
  - Doesn’t take into account time as a factor

Implementation

- More testing and tuning
  - Confidence Interval on RMSE and using bootstrapped data
  - More scale up data -> timing of 10 M set, compare everything in clusters.
  - More gridsearching and optimization
- Small scale roll out: A/B Testing
  - We can dip our feet in, before we jump
  - We can better analyze scale up timing, RMSE, memory, cost etc.
- If we don't switch at a minimum we should investigate weighing each mean (user, movie, global) in the mean of means to determine if we can improve our current algorithm.
- Take into account ratings with respect to time