Recommendation System: amazon Data Analysis

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Advisor: Julian McAuley
Team: Nirmal Budhathoki, Chris Chen, Christopher "Toby" Moreno, Nolan Thomas

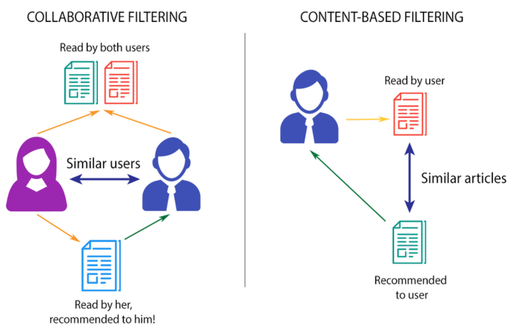
Masters of Advanced Study in Data Science and Engineering

Introduction

A ***recommender system*** can provide product recommendation to the customers based on their past history of purchases, reviews, and/or product searches. There are two basic architectures for a recommender system:

A. Content-based filtering: Recommend items that are similar to those that a customer rated high in the past. User profile and item profile are created to capture unique characteristics of the user and the item which is used to predict the heuristics by computing the similarity scores between the user's and item's vectors.

B. Collaborative filtering: It relies on past user behavior like previous products reviews or ratings to infer new user-item associations. The key advantage of this approach is that it does not require user profile or item profile. The drawback of this approach is that it suffers from what is known as the *cold start* problem. The model performance heavily depends on the 'density' of user-item interaction. When a new user or a new item come into the system, the model prediction deteriorate.



Collaborative filtering can broadly fall under either neighborhood-based similarity models or latent-factor models.

Bayesian Personalized Ranking (BPR):

The algorithm addresses the One Class Collaborative Filtering (OCCF) problem by turning it into a ranking problem and implicitly assuming that users prefer items that they have already interacted. Instead of applying rating prediction techniques, BPR ranks candidate items for a user without calculating a 'virtual' rating.

Problem Statement

To build a recommender system that is capable of capturing customer's preferences by:

i. *Exploring other relevant features besides the latent factors.*
ii. *Analyzing how price sensitivity affects the user's purchase behavior.*

One of the objectives is also to learn how the sparsity of data affects feature(s).

Analysis Workflow

The analysis workflow consists of acquiring data, preparing data, analyzing data (which involves modeling), reporting the results and taking actions based on the analysis, and positive or negative feedbacks from real time users.

The data is **acquired** from http://jmcauley.ucsd.edu/data/amazon/ , owned by Professor Julian McAuley. This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. Focused categories: *Women's Clothing* and *Electronics*.

The **data preparation** phase consists of: **exploring and pre-processing data**. Some of the findings from exploration are people mostly tend to give higher ratings in amazon, the price attribute is positively skewed (many items fall under low price ranges), the log transformation improves the distribution to normal, but is that sufficient to be proven as a relevant feature? This is yet to be discovered in analysis phase. **Feature engineering** is also a part of our data preprocessing to make the input data ready for analysis. Some of the features generated and implemented for clothing category (in this example) are:

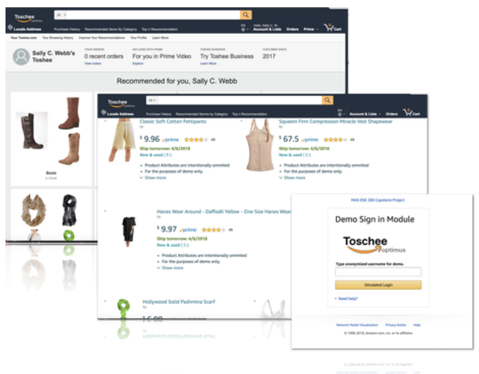
Category Tags	Product exists across multiple categories. Those subcategories are turned into a set of 697 tags	Very Beneficial
Subcategory (Level 4)	70 sub-categories at Level-4 in the hierarchical classification of the product	Beneficial
Sentiment/Polarity	Using Text Blob sentiment polarity	No value add
Seasonal	Meteorological seasons	No value add
Brand	1769 brands	Beneficial
Overall Rating	Simple encoding	No value add
Price	Log transformation	Beneficial
Bi-gram words	Most frequently used 4525 bi-grams from product title	Beneficial

After data cleanup and generating the features, the next phase is **analysis or modeling**. Our analysis method is Classification with ranking using BPR. Any machine learning (ML) algorithm can be broken down as:

ML algorithm = Model + Loss function + Optimization

The model we have used is Matrix Factorization, the loss function being used is BPR, and the optimization method used is Stochastic Gradient Descent. Area Under Curve (AUC) is used as a metric to measure the accuracy. This step is repetitive allowing reverse feeding the model with results to improve performance.

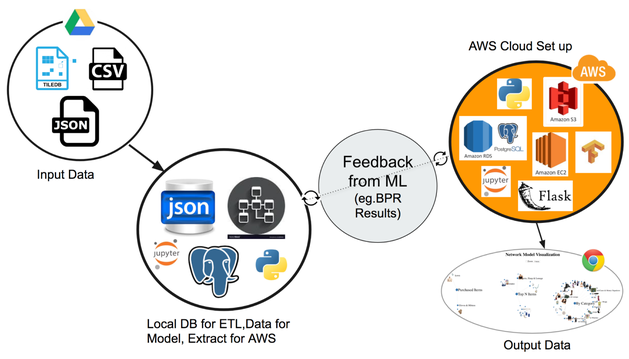
For **reporting,** a web-based tool is designed using JavaScript, Python, and Flask library to present the recommended items for the sampled test users.



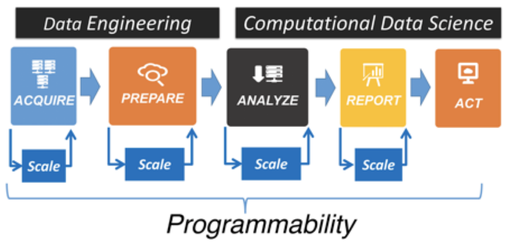
For **Act,** our model is not yet implemented to collect feedbacks from real time users, whether they liked and purchased the recommended items or they skipped or marked them as not useful.

Solution Architecture

The input data comes in various formats like json, csv etc., and are stored in G-Drive. Local database in PostgreSQL is set up for ETL, and feature engineering. All the code is written to test in smaller iteration in local machine, and the model is ran with higher epochs, tested, optimized, and evaluated in AWS cloud. The output data is presented in a webtool.



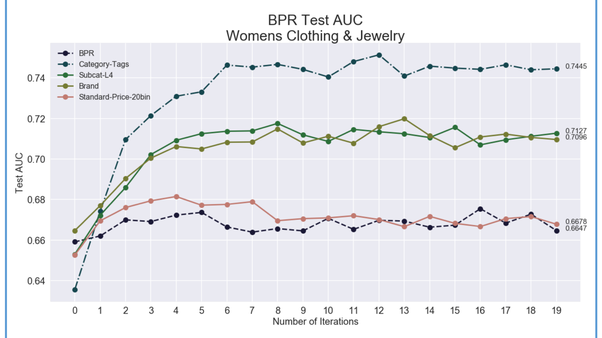
Scalability

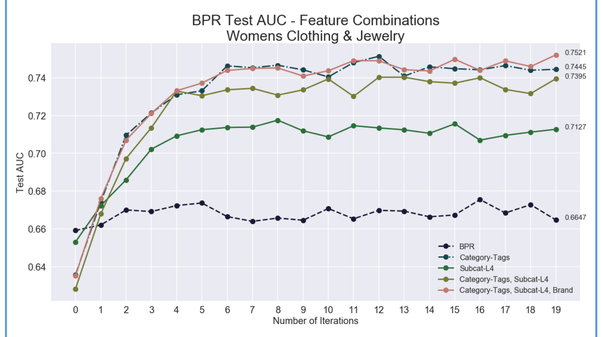


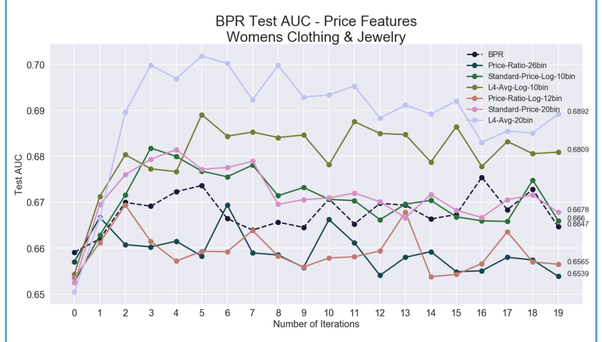
Scalability matters in every steps of our analysis process. The data pipeline is designed to read, parse, load data, and perform feature engineering. Modeling is based on TensorFlow, which is highly scalable across multiple CPUs or GPUs to distribute the workload. Matrix Factorization approach incorporates latent features as well as additional item features into predictors of user's opinion while scaling to larger datasets.

Key Insights/ Results

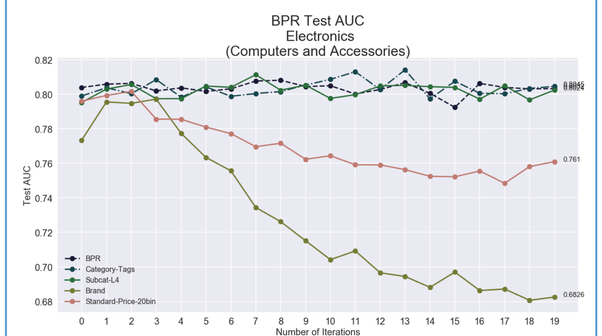
Women's Clothing Category

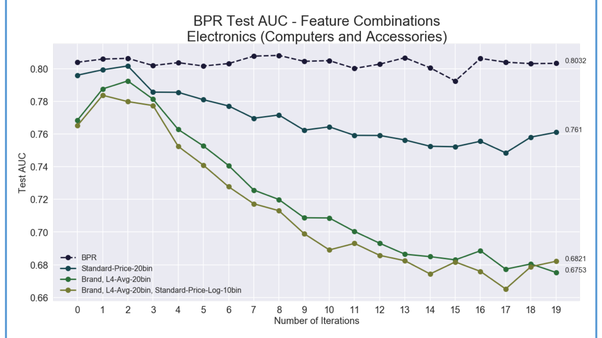


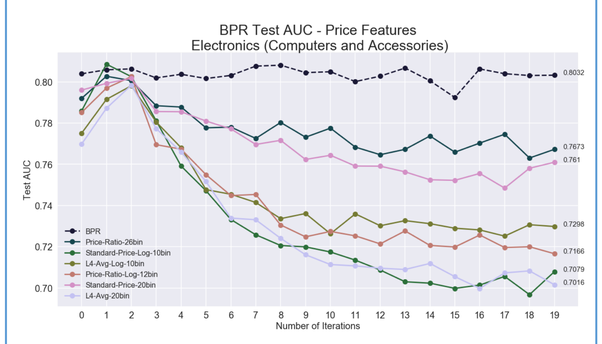




Electronics Category







Acknowledgements

We would like to pass our big thanks to:

- Julian McAuley, Project Advisor
- Ilkay Altintas, Project Coordinator/ Lecturer
- Sharknado: Cohort 2's Amazon Project Group
- All Faculty members of DSE program
- All Administrative staffs of DSE program
- All our classmates from Cohort 3

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com