Measuring the role of visual features and marketing bias in product recommendations

Shubhanshu Gupta gupta.shubh@northeastern.edu
Kaushik Nishtala nishtala.k@northeastern.edu
Kathan Patel patel.kat@northeastern.edu

1 PROBLEM DESCRIPTION

Recommender systems play a major role in allowing users to discover items "similar" to their personal preference amongst huge corpora of products in a commercial setting. Our goal is to build an optimal product recommender system using methods that rely on content-based filtering, where in we group products together based on the similarity of their attributes, as well as collaborative filtering which allows us to cluster similar users to predict their product ratings and preferences. Having built such a modeling framework, we would then channel our focus towards the impact of product images on the model's efficacy to produce better recommendations by employing Convolutional Neural Networks, as it would be interesting to use their high compute power to glean additional information from images to improve model performance. Computationally, this follows a series of methods with which we predict users' ratings on the products they have not interacted with yet, and surface a ranked list of top k products they would most likely be interested in. We would then apply an appropriate evaluation criteria such as RMSE, Normalized Discounted Cumulative Gain etc., to rate the performance of our model.

Additionally, a huge part of our work deals with the dynamics of consumer-product interactions. However, these consumer preference patterns can often be influenced by an inherent bias in the way the product is marketed, for instance due to the selection of a particular human model in a product image. This bias effects the model by surfacing irrelevant recommendations to users underrepresented in the interaction data. As an example, when a particular gym equipment is promoted using a stereotypically 'male' image, a female consumer who would potentially be interested in it may be less likely to interact with it. To that end, we plan to tackle this interesting phenomenon by implementing common collaborative filtering algorithms to identify and address the bias by studying any notable deviations of the resulting model outputs/errors from the interaction data.

2 ALGORITHMS

For content-based filtering, we find the similarity of the product attributes (textual and images) to estimate the ITEM x ITEM affinity. We use methods such as TF-IDF, Nearest Neighbours, Bag of Words or Word2Vec in order to featurize textual data based on semantic similarity. For product images, we employ a visual product similarity with the help of deep neural networks to featurize the edges, shapes, parts of an image. This is inspired by E. H. Ahmed et al who validated the improvement in Neural Network performance by adding

images to their text feature-only model. Likewise for collaborative filtering, we use matrix factorization technique among others to help provide more accurate and scalable model outputs than standard clustering due to the existence of latent features in the model.

3 Dataset

We make use of two separate Amazon datasets in our study. First Dataset[1] contains information regarding Amazon products and reviews. Our main interest lies in leveraging features such as user reviews, ratings etc. for the text-based collaborative recommendation system. Furthermore, we utilize features related to product attributes such as price, brand, size, specifications etc., and the images of the product for the content-based recommendation system. On the other hand, the second dataset[2] includes additional details about user genders and identities that are critical for our study to identify marketing bias. As the data is feature rich ranging from textual data and numerical data to images, we plan to use several pre-processing techniques to clean up the data. For textual data, we intend to use techniques such as stemming, lemmatization, text normalization etc. to extract critical information and to transform them into model-ready format. Additionally, we would like to scale all images to the same size before feeding them as input to our neural network. We may also perform other pre-processing techniques such as taking care of missing data, scaling of features, dimensionality reduction, merging datasets etc., as and when the need arises.

4 Libraries and tools

- Some of the libraries that we intend to make use of include Python Data Stack (numpy, pandas, scipy, matplotlib etc), seaborn, random, collections, requests, BeautifulSoup, nltk, sklearn (CountVectorizer, TfidfVectorizer, cosine_similarity), Tensorflow/Keras, Surprise[3] , crab[4] among others.
- Tools/IDE we plan to use include Google Colab, Jupyter Notebooks and/or Spyder IDE.

5 Results

Ideally, with our optimal recommender model, we hope to find an improved model performance with the addition of images to the feature set. Furthermore, We also expect to identify systematic deviations across several consumer-product categories in terms of rating prediction error, and considerable deviations of the resulting model outputs from the interaction data. If we fail to reach a desirable outcome, we intend to employ a backtrack-and-iterate approach and experiment with several parameters (such as varying number of latent features, regularization and initialization effects etc) via Cross Validation in order to find the optimal model that steers us towards the desired result. Finally, we hope to deep dive into this study by practicing a clear division of labor into several segments - EDA (all) + content-based + collaborative-based + marketing bias, thereby allowing us to collaborate efficiently and put together a coherent study.

[1] Amazon Review Data (2018) Source
[2] Marketing Bias Data Source
[3] Surprise Library
[4] Crab Framework