Applied ML Final Project Proposal - Group 29

Zhining Qiu, Zhenyu Yuan, Ran Pan, Tao Wang, and Smaranjit Ghose

Background and context

With the rapid growth of the Internet, more and more famous retail brands begin to start their own online stores and E-commerce sites. However, with an extensive selection of products and too many choices, users often have trouble finding what really interests them or what they are looking for, and eventually fail to make purchases. Therefore, designing an efficient recommendation system is imperative for every E-commerce website to present their customers with high quality information and products that are relevant to their personal preferences. More importantly, a successful online recommendation system can increase customer retention rates and reduce product returns, which will benefit the E-commerce site in the long term.

In this project, we would like to design and implement a product recommendation system for H&M Group, a multinational clothing company that provides fast-fashion clothing for people in all age groups, in order to help their customers to locate their products of interest and achieve customer satisfaction.

Our goal is to provide top 10 articles each customer is most likely to buy based on their previous shopping experience. Customers can get what they possibly want and clerks can easily find recommendations for valued individuals.

Data

The dataset is the transactions of customers and the metadata of customers and products. The transactions are represented as tuples of user id and product id, which means that a user made a transaction for a specific product.

The available meta data spans from simple data, such as garment type and customer age, to text data like product descriptions, to image data from garment images. Specifically, we have image, id, name, type, group, graphical appearance, color and department features for the product. And there is id, FN, active information, club member status, fashion news frequency age and postal code features.

Overview of methodology

We plan to convert various features of a product into a fixed length feature embedding, and then fit an autoregressive model on each customer's purchase activities. For prediction, we'll feed the trained autoregressive model with a customer's history, and take it's output as our prediction. Basically, this method is analogous to Seq-to-Seq methods in NLP.

Product embedding

- For categorical features like colors, we'll map each value into an n-dimensional vector, and concat the columns together
- For the name of the product and the product description, we'll apply some existing NLP model (maybe some variant of BERT) to obtain a representation
- All the above vectors will be concatenated into a long vector embedding for each product, and all the embedding layers will be trainable

Autoregressor

We intend to train an autoregressor (e.g. LSTM) on each user's purchase history. The input to this model would be the embedding obtained above and the training process would be analogous to the language model task in NLP.