Master thesis Jaime

Jaime Ferrando Huertas

March 2021

1 Introduction

This documents aims to provide a brief introduction on the HM dataset that will be released on Kaggle as well as some ideas for a potential master thesis working with this dataset. The dataset will be released with a Kaggle competition in the future 2021 but the master thesis work can already start before that as we already have access to all the data. Having early access to the data also means we could identify flaws or improvements on the dataset and if we were to need more data or extra features that were not included in the original dataset that could be solved an added to the prerelease dataset.

The purpose of this competition is to create a ranked list of products for each user. This list represents the recommendations from the system to the user and is yet not be defined how to evaluate this recommendations, some evaluating examples: evaluate on every last customer purchase, evaluate in a 7 window period of purchases (one customer can produce multiple purchases, or none).

2 Dataset

For this dataset there are multiple tables but for the sake of simplicity i provide a condensed explanation of the data contained across all tables by grouping in 3 main tables.

- **List of transactions**: Contains article_id, date of transaction, customer_id and price for the item. We have 2 years of transaction data and the possibility to extend that further in the past if necessary.
- **Article data**: For each unique article_id in transaction data we have a set of features, some of them being: article name, clothing category (shorts, shirt, etc), color, graphical appearance(solid, stripes, pattern, ...), section (kids, young girl, ladies trend, ...), picture and a text description.
- **Customer data**: For each customer_id in transaction data we have their age, postal_code, active_status, and frequency on reading fashion news.

We have a history of for every customer i along with a set of features for every article j and our recommendation work should try to predict what next article j our customers will buy.

3 Ideas

3.1 Space representation and similarity

A classical approach is to use the features of the articles in a customer purchase history $\mathcal{X}_i$ to create a user representation. This representation will work as a user profile and we can later use it to find the closest article j in space and serve those as recommendations.

When article features are not representative enough we can create a D dimensional representation of our articles by using embeddings of size D. To learn this we base on word2vec and treat user purchase history as a sentence and learn article embedding representation from articles bought before/after this one.

3.2 History as a context to predict next event

Basing on google work for youtube recommendations [1] we could try something similar on our users behaviour. To do so we start on the purchase history $\mathcal{X}_i$ for every customer i with the corresponding timestamp t of each purchase. We can denote all purchases a customer has made at time t as:

$$\mathcal{X}_{i,t} = \{e = (i, j, t) \in \mathcal{E} \mid e_0 = i \wedge e_3 < t\} \subset \mathcal{X}_i \tag{1}$$

We now try to predict the next article j the customer will buy at time t based on his history: $Pr\left(j \mid i, t, X_{i,t}\right)$. Google work [1] does not include video data in this approach and lets the rnn figure learn the video representation as embeddings.

This line of research has been continued by other teams by adding a hierarchical context [3] and even adding attention layers to optimize it [5].

3.3 Deep learning alternatives to collaborative filtering

There have been multiple deep learning approaches to mimic collaborative filtering, Deep ICF [4] being one of earliest and most famous ones. The state of the art for deep learning models that mimic collaborative filtering is the DRLM [2] released by facebook. Deep ICF created embeddings for both user and items to later feed their pairwise interaction to a multilayer perceptron. DLRM adds a new layer of complexity creating embeddings an items categorical features rather than the items itself, DLRM splits numerical and categorical features and computes the pairwise interaction of numerical features and categorical embeddings representation. DRLM also removes the user embedding and adds its features into the list of numeric/categorical features. This approach also allows for a more generalist approach as it can receive a new item and create a

good representation of it basing on its features, something that Deep ICF could not do.

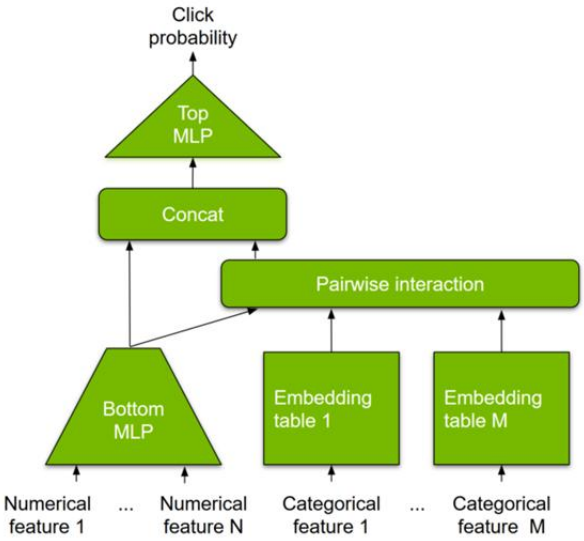

Figure 1: DLRM architectureα

4 Conclusion

With H&M being a fastfashion company and the short-timed nature of trends we believe there is potential to create recommendations using both time limited features (such as user history) and stationary features(numeric and categorical) such as the ones we have in our articles data. We present three ideas on how this features have been dealt in the past to create recommendation systems, let them be used as inspiration to create one that fits our needs.

References

[1] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed Chi. Latent cross: Making use of context in recurrent recommender systems. pages 46–54, 02 2018.

[2] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. Deep

learning recommendation model for personalization and recommendation systems. *CoRR*, abs/1906.00091, 2019.

[3] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. Personalizing session-based recommendations with hierarchical recurrent neural networks. pages 130–137, 08 2017.

[4] Feng Xue, Xiangnan He, Xiang Wang, Jiandong Xu, Kai Liu, and Richang Hong. Deep item-based collaborative filtering for top-n recommendation. *CoRR*, abs/1811.04392, 2018.

[5] Haochao Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. Sequential recommender system based on hierarchical attention networks. pages 3926–3932, 07 2018.