Computing Non-negative Rank

Abstract

1. Introduction

Understanding customers is crucial to Amazon's business. This includes identifying customers' interests, purchase patterns and demographics to improve customer experience. Amazon has a huge customer base with rich amount of data on customer purchases, subscriptions, product reviews, etc. However it has limited access to the customer demographics, e.g. age, gender, income, etc. Recently there has been interest on inferring the demographic informations for the account holder. However, as multiple persons usually use the same account to buy products, inferring just the account holder's demographics seems to be sub-optimal. In this work, we intend to discover the demographics of all persons who are using the same account. For simplicity, we focus on discovering age range of each member in the household. The six age ranges in consideration are : 18-24, 25-34, 35-44, 45-54, 55-64 and above 65. Knowing the age composition of a complete household helps to improve the quality of recommendation and targeting, e.g. knowing that there is a baby (or young adult) in a household helps targeting for diaper (or back to college) subscriptions.

There are multiple *challenges* involved in discovering age composition of a household.

- *Low determinism:* The fraction of purchases deterministic to some demographic age group is very small e.g. the number of products specific to people older than 65 is significantly low.
- *Noise in purchases:* Significant number of purchases from an account are for members outside the household. This has been observed from an analysis based on the household data (obtained by Kindle) [1].
- *Label Noise:* To the best of our knowledge, there is no dataset inside Amazon containing the complete age composition of households. The household data obtained by the Kindle team is partially labelled. Also, the account holder's age present in Experian is 63% accurate [1].

These make the learning of a supervised model to predict household age composition more challenging. In this paper we address these challenges by making the following contributions.

- To address the issue of *low determinism* we don't use the purchased ASINs directly in our modeling. Rather we obtain the title and brand names of the ASINs purchased (along with other features) and process them through a mutual information based feature selection to identify title and brand words most deterministic of age bands.
- It is intuitive that the distribution of household compositions is not uniform i.e. certain household compositions are more frequent than others e.g. [25-34, 35-44] is more likely than [18-24, 35-44] . This hints that the labels are considerably correlated and to model the correlation we explore Conditional Bernoulli mixture model (CBM) which is a multi-label classification algorithm handling label correlations.
- Due to the unavailability of high quality household data, we train our models on account holder age information of single households. To handle the noise in labels we use models that are robust to label noise. We demonstrate the models robustness via thoroughly designed experiments. Further, we experiment with various matching criteria used to match the data from Experian with Amazon customers, to understand whether the label noise is a result of the matching operation.
- To investigate the effect of noisy purchases, we experiment with low rank representations (SVD) and provide useful insights.
- To demonstrate the efficacy of our models we conducted rigorous experiments in predicting the age and gender composition and present applicable results.

[1] To see more about the dataset please see section **??**

For each account, our model also predicts the account holder age group. This set of prediction is in production through Bulls-eye platform and used by more than 50 different teams across Amazon.

In section **??**, we discuss different approaches in modeling the household prediction including algorithms handling label correlations. We present various types of experiments to show the efficacy of the models in predicting household age composition with a thorough analysis of the predictions in section **??**. A brief concluding remarks with potential future directions is presented in section **??**.

References

Christopher, D Manning, Prabhakar, Raghavan, and Hinrich, SCHÜTZE. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151:177, 2008.

Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dembczyński, Krzysztof, Waegeman, Willem, and Hüllermeier, Eyke. Joint mode estimation in multi-label classification by chaining. In *CoLISD 2011 (Collective Learning and Inference on Structured Data 2011)*, 2011.

Freund, Yoav, Schapire, Robert E, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pp. 148–156, 1996.

Koh, Pang Wei and Liang, Percy. Understanding black-box predictions via influence functions. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Kumar, Abhishek, Vembu, Shankar, Menon, Aditya Krishna, and Elkan, Charles. Learning and inference in probabilistic classifier chains with beam search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 665–680. Springer, 2012.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, Cheng, Wang, Bingyu, Pavlu, Virgil, and Aslam, Javed. Conditional bernoulli mixtures for multi-label classification. In *International Conference on Machine Learning*, pp. 2482–2491, 2016.

Liu, Weiwei and Tsang, Ivor. On the optimality of classifier chain for multi-label classification. In *Advances in Neural Information Processing Systems*, pp. 712–720, 2015.

Mcauliffe, Jon D and Blei, David M. Supervised topic models. In *Advances in neural information processing systems*, 2008.

Pani, Jagdeep, A, Pooja, Majumder, Anirban, Chaoji, Vineet, Khare, Vineet, and Rastogi, Rajeev. A generative model to learn household composition of customers. In *Amazon Machine Learning Conference*, 2016.

Read, Jesse, Pfahringer, Bernhard, Holmes, Geoff, and Frank, Eibe. Classifier chains for multi-label classification. *Machine Learning and Knowledge Discovery in Databases*, pp. 254–269, 2009.

Sohoney, Saurabh, Roy, Gourav, Kaveri, Sivaramakrishnan, Khare, Vineet, and Chaoji, Vineet. Scalable clustering for mining interest circles from customer behavior at amazon. In *Amazon Machine Learning Conference*, 2016.

Tsoumakas, Grigorios and Katakis, Ioannis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 2006.