Parameter-efficient multi-modal embeddings for representing diverse user interests

Maciej Kula
maciej.kula@gmail.com

ABSTRACT

Most existing recommendation approaches implicitly treat user tastes as unimodal, resulting in an average-of-tastes representations when multiple distinct interests are present. We show that appropriately modelling the multi-faceted nature of user tastes through a mixture-of-tastes model leads to large increases in recommendation quality at a very modest cost in model complexity. Our result holds both for deep sequence-based and traditional factorization models.

KEYWORDS

Recommender Systems, Matrix Factorization, Binary Representations

1 INTRODUCTION

2 EXPERIMENTS

Datasets

For our experiments, we use the following datasets.

Movielens 10M. A dataset of 10 million ratings across 10,000 movies and 72,000 users [1].

Goodbooks. A dataset of 6 million ratings across 53,000 users and 10,000 most popular books from the Goodbooks online book recommendation and sharing service [4].

Amazon. A dataset of ratings and reviews gathered from the Amazon online shopping service. After pruning users and items with fewer than 10 ratings, the dataset contains approximately 4 million ratings from 100,000 users over 114,000 items.

We treat all datasets as implicit feedback datasets, where the existence of an edge between a user and an item expresses implicit preference, and the lack of an edge implicit lack of preference.

Experimental setup

For factorization models, we split the interaction datasets randomly into train, validation, and test sets. We use 80% of interactions for training, and 10% each for validation and testing. We make no effort to ensure that all items and users in the validation and test sets have a minimum number of training interactions. Our results therefore represent partial cold-start conditions.

For sequence-based models, we order all interactions chronologically, and split the dataset by randomly assigning users into train, validation, and test sets. This means that the train, test, and validation sets are disjoint along the user dimension. For each dataset, we define a maximum interaction sequence length. This is set to 100 for the Goodbooks and Movielens datasets, and 50 for the Amazon dataset, as the interaction sequences in the Amazon dataset are generally shorter. Sequences shorter than the maximum sequence length are padded with zeros.

We use mean reciprocal rank (MRR) as our measure of model quality. In factorization models, we use the user representations obtained from the training set to construct rankings over items in the test set. In sequence models, we use the last element of the test interaction sequence as the prediction target; the remaining elements are used to compute the user representation.

We perform extensive hyperparameter optimization across both our proposed models and all baselines to mitigate the danger that our results driven either by models' sensitivity to hyperparameter choices (when a superior model performs poorly because of a poor choice of hyperparameters), or experimenter bias (in favouring the optimization of the novel model). We optimize batch size, number of training epochs, the learning rate, L2 regularization weight, the loss function, and (where appropriate) the number of taste mixture components.

We experiment with two loss functions:

- Bayesian personalised ranking (BPR, Rendle et al. [2]), and
- adaptive sampling maximum margin loss, following Weston et al. [3].

For both loss functions, for any known positive user-item interaction pair (u, i), we uniformly sample an implicit negative item $j \in S^-$. For BPR, the loss for any such triplet is given by

$$1 - \sigma\left(r_{ui} - r_{uj}\right), \tag{1}$$

where σ denotes the sigmoid function. The adaptive sampling loss is given by

$$\left|1 - r_{ui} + r_{uj}\right|_+ . \tag{2}$$

For any (u, i) pair, if the sampled negative item j results in a zero loss (that is, the desired pairwise ordering is not violated), a new negative item is sampled, up to a total of k attempts. This leads the model to perform more gradient updates in areas where its ranking performance is poorest.

Across all of our experiments, the adaptive maximum margin loss consistently outperforms the BPR loss on both baseline and mixture models. We therefore only report results for the adaptive loss.

REFERENCES

[1] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2016), 19.

[2] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[3] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, Vol. 11. 2764–2770.

[4] Zygmunt Zajac. 2017. Goodbooks-10K dataset. https://github.com/zygmuntz/goodbooks-10k. *GitHub* (2017).

Table 1: Experimental results

Model	Movielens 10M	Amazon	Goodbooks-10K
LSTM	0.09	0.15	0.116
Mixture-LSTM	0.1	0.189	0.136

[1] Ratio of binary model MRR to real-valued model MRR

[2] Predictions Per Millisecond: how many items can be scored per millisecond

[3] Ratio of binary PPMS to real-valued PPMS

[4] Ratio of memory required to store binary vs. real-valued parameters