TC1 Project: Know your customers

Aris Tritas Laurent Cetinsoy

November 14, 2016

Abstract

We are given a training set describing the ratings assigned by customers to items (e.g a firm's services). A collaborating filter problem is addressed with the following methods: NMF approach and autoencoders.

1 Introduction

In collaborative filtering, each item i of a collection is rated by one or more members of a population n_u. The aim is to predict the rating that a person would give to an unseen item, so as to be able to recommend new items to users such that it fits their preferences with high probability.

In this setup, let R represent a $n \times m$ matrix. The rating of the j^{th} item from the i^{th} user is r_{ij}.

1.1 Data

Dataset size: 2536705 examples in the training set, 1306637 in the test set. Total number of users 93705 (of which only 92088 are in the training set) and 3561 items. The dataset is very sparse, less than 1% of the ratings matrix is filled. Ratings distribution: [114214, 251596, 728518, 860463, 581913] In proportion of the total number of ratings: [0.0450, 0.0991, 0.2872, 0.3392, 0.2293]

1.2 Hidden Markov Model

It is possible to model the problem of sequential prediction using a hidden Markov model. In order to do that efficiently, it is assumed that

1.3 Non negative matrix factorization

In non negative factorizations we seek SPD matrices U and V such that the rating product matrix can be factorized as $R = UV^T$. It is possible to find such matrices that minimize the reconstruction error defined below by using an iterative optimization method (e.g gradient descent):

$$\begin{aligned}\arg\min_{U,V} &\frac{1}{2}||R - UV||^2 \\ &+\alpha\lambda_1(||U||_1 + ||V||_1) \\ &+\frac{1}{2}\alpha\lambda_2(||U||^2 + ||V||^2)\end{aligned}$$

The reconstruction error is computed over all non negative coefficients of the R matrix. In other words, the algorithm only tries to recover the available ratings.

1.4 ALS-WR

The Alternating Least Squares with Regularization algorithm [2] differs from NMF for two main reasons: a regularization term and the fact that the minimization is done by alternatively updating U while keeping V fixed and then updating V while keeping U fixed.

The parameters used were $\lambda = 0.01$ and a latent dimension $d = 100$.

1.5 Auto-encoders

The auto-encoder is a neural network approach used in unsupervised learning for dimensionality reduction. It aims at lowering the reconstruction error, which can be written as:

$$||X - NN(X)||$$

.

In the supervised classification task, the auto-encoder's objective is to lower a prediction error over a training set.

The approach proposed by [1] was used to tackle our collaborating filter problem. Two different auto-encoders can be trained depending on whether the users or the items are considered as training cases.

2 Result

The metric used to evaluate the performance of a training method for this type of problem is either the RMSE (root mean squared error) or the MAE (mean absolute error). The scores below were computed on a validation set selected uniformly at random amongst the training set:

NMF	ALS-WR	User AE	Item AE
3.70	0.934	0.913	0.938

3 Conclusion

[1] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber and Rong Pan - Large-scale Parallel Collaborative Filtering for the Netflix Prize.

References

[1] Florian Strub, Jeremie Mary, and Romaric Gaudel. Hybrid collaborative filtering with autoencoders. *arXiv preprint arXiv:1603.00806*, 2016.

[2] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *International Conference on Algorithmic Applications in Management*, pages 337–348. Springer, 2008.