CONJUGATE-PRIOR-REGULARIZED MULTINOMIAL PLSA FOR COLLABORATIVE FILTERING

Marcus Klasson

Dept. of Mathematical Sciences, Lund University, Sweden

ABSTRACT

When making predictions on unseen data with *probabilistic Latent Semantic Analysis* (pLSA) in a collaborative filtering setup the overfitting is a severe problem. By applying a conjugate prior regularized approach to the multinomial pLSA, the aim is to provide more robust learning on discrete rating data in the commonly sparse data sets. In this article, the proposed regularization method is applied to counteract the overfitting problem and therefore reducing the prediction errors.

1. INTRODUCTION

Collaborative filtering (CF) is an approach within recommender systems where the recommendations are based which items that consumers with similar tastes have purchased or liked in the past [1, 2]. This paper considers the model-based CF method called *probabilistic Latent Semantic Analysis* (pLSA), which has many applications in information retrieval and filtering and is an important tool machine learning in text. The three main challenges with pLSA in CF is overfitting [1–3], which results in less reliable model parameters and increases the prediction errors for unseen data, the sparseness in the data, i.e. the small amount of ratings that are actually provided to the system, and also the differences in how users utilizes the rating scale. In order to mitigate the overfitting and sparsity problem, we propose applying *Conjugate-Prior-Regularization* to the pLSA model [4–6]. The resulting model applies *maximum a posteriori* (MAP) estimation with conjugate prior distributions, which is an extension of *Expectation Maximization* (EM) algorithm.

The following notation is used: we define a set of users $\mathcal{U} = \{u_1, \dots, u_m\}$ and items $\mathcal{I} = \{i_1, \dots, i_n\}$. The users have the opportunity to rate items with a preference value from an explicit rating scale $\mathcal{R}$, where the given rating data $r_{u,i}$ is stored in an $m \times n$ matrix called $\mathbf{R}$. Latent states $z \in \{z_1, \dots, z_K\}$ are introduced to the model, where K is the amount of possible states. The setup assumes also that the *forced prediction* is used, which mimics when items are presented as recommendations to a user and one is interested in foreseeing the user's response [1, 2]. The non-regularized model is thus given by the mixture model [2]

$$P(r|u, i; \theta) = \sum_z P(r|i, z)P(z|u), \tag{1}$$

where the parameter vector $\theta = \{P(r|i, z), P(z|u)\}$ and contains every probability distribution in the model and also that $P(z|u)$ and $P(r|i, z)$ are proper conditioned probabilities, i.e. $\sum_z P(z|u) = 1$ and $\sum_r P(r|i, z) = 1$. Unlike probabilistic user-clustering models, where each user would be associated to a single latent state, every single observation $\langle u, i, r\rangle$ is connected to a latent state in pLSA [1, 2].

2. MAP ESTIMATION WITH CONJUGATE PRIORS

The conjugate prior of a multinomial distribtuion is the Dirichlet distribution and using such a prior means that the posterior distribution, or the product of the likelihood and the prior distribution, also will be Dirichlet distributed [4, 5]. Assuming that $P(r|i, z)$ and $P(z|u)$ are independent, the prior distribution is expressed by [6]

$$\begin{aligned} P(\theta) &= \text{Dir}(\{P(r|i, z)\}|\{\gamma_{i,r,z}\}) \cdot \text{Dir}(\{P(z|u)\}|\{\gamma_{u,z}\}) \\ &\propto \prod_z \left[\prod_{i,r} P(r|i, z)^{\gamma_{i,r,z}-1} \prod_u P(z|u)^{\gamma_{u,z}-1} \right], \end{aligned} \tag{2}$$

where $\varphi = \{\gamma_{i,r,z}, \gamma_{u,z}\}$ are the respective hyperparameters to the conjugate prior distributions. The hyperparameters control the degree of regularization to their corresponding parameter by penalizing parameter distributions that have overfitted towards a certain value. When increasing the hyperparameter value more penalization is added to the estimates, where the outcome is a more equally distributed parameters.

For estimation of θ, *variational probabilities* $Q(z|u, i, r; \theta)$ is introduced to the log-likelihood function that are supposed to model the probability for an observation to be associated with state z. The MAP-based EM algorithm is then alternated between the expectation and maximization step until the parameter values have converged. In the E-step, the optimal variational probabilities, denoted by Q^*, are estimated as [2, 4, 5]

$$Q^*(z|u, i, r; \theta) = \frac{P(r|i, z)P(z|u)}{\sum z'(P(r|i, z')P(z'|u)}. \tag{3}$$

Thereafter in the M-step, the new regularized parameter estimates are computed with the Q^*-distributions with the follo-

wing expressions

$$P(z|u) = \frac{\sum_{\langle u',i,r\rangle:u'=u} Q^*(z|u,i,r;\theta) + (\gamma_{u,z} - 1)}{\sum_{z'}\sum_{\langle u',i,r\rangle:u'=u} Q^*(z'|u,i,r;\theta) + (\gamma_{u,z'} - 1)} \tag{4a}$$

$$P(r|i,z) = \frac{\sum_{\langle u,i',r'\rangle:i'=i,\, r'=r} Q^*(z|u,i,r;\theta) + (\gamma_{i,r,z} - 1)}{\sum_{\langle u,i',r\rangle:i'=i} Q^*(z|u,i,r;\theta) + (\gamma_{i,r,z} - 1)}, \tag{4b}$$

where the prime signs under the summations denote a fixed variable for the conditional probability computed. Note that this result reduces to the EM-algorithm presented in [2] if all hyperparameters are set to 1.

3. EXPERIMENTAL RESULTS

To investigate the proposed conjugate-prior-regularized MAP pLSAs capability to mitigate overfitting, the proposed model is compared with a pLSA using the standard EM algorithm with an early stopping (ES) condition and also the so called Pop item-average estimator. This was examined with the EachMovie data set, which contains $2,811,718$ ratings entered by $61,265$ users and 1623 items (movies) [10]. The data was randomly divided into three sets; training, validation, and test data sets. To reduce the variance, this subdivision was repeated ten times on the original data set. The evaluation is made by computing the prediction errors for the test set with

$$\text{RMSE} = \sqrt{\frac{1}{|\mathbf{R}|}\sum_{\langle u,i\rangle\in\mathbf{R}} (r_{u,i} - \hat{r}_{u,i})^2}$$

$$\text{MAE} = \frac{1}{|\mathbf{R}|}\sum_{\langle u,i\rangle\in\mathbf{R}} |r_{u,i} - \hat{r}_{u,i}|,$$

where $\hat{r}_{u,i}$ is the predicted rating to the actual rating $r_{u,i}$ and given by

$$\text{E}[r|u,i] = \sum_{r\in\mathcal{R}} r\, P(r|u,i) = \sum_{r\in\mathcal{R}} r \sum_{z} P(r|i,z)P(z|u). \tag{5}$$

A cross-validation method was used for finding suitable hyperparameters, where the MAP pLSA parameters are trained with different hyperparameter value combinations and evaluated on the test sets RMSE. For the ES pLSA it was found that the lowest prediction errors were recieved with the latent state size $K = 200$. The conjugate-prior-regularized MAP pLSA performed best with $K = 50$ and hyperparmeter values $\gamma_{u,z} = 1.08$ and $\gamma_{i,r,z} = 1.5$. The smallest prediction errors were achieved with the 10th picked data sets and models training procedures are shown in figure 1. It is clear that the proposed MAP pLSA has mitigated the overfitting, since the gap between its training and test errors is reduced compared to the ES pLSA. The means and standard deviations of RMSE and MAE for both pLSA models and the pop estimators are shown in table 1.

		Proposed	ES pLSA	Pop
RMSE	mean	1.2375	1.2727	1.3712
	std	0.0064	0.0070	0.0062
MAE	mean	0.9711	0.9834	1.0908
	std	0.0047	0.0052	0.0048

Table 1. Prediction error means and standard deviations from the EachMovie data set.

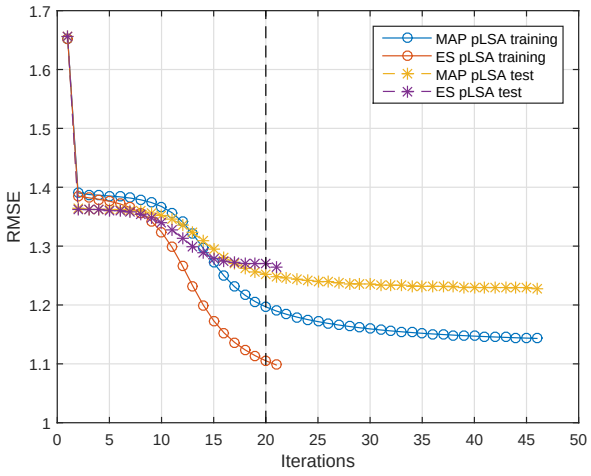

Fig. 1. RMSE over iterations for MAP and ES pLSA. The graph hows the RMSE for both training and test sets and also when the early stopping occurs as the dashed line.

4. CONCLUSIONS AND FURTHER WORK

It was found that hyperparameter $\gamma_{u,z}$ has a greater impact on decreasing the prediction error than $\gamma_{i,r,z}$ for the EachMovie data. Also, the proposed conjugate-prior-regularized pLSA overcomes the ES pLSA in both reducing the overfitting and prediction errors. Since model complexity is reduced to $K = 50$ in the proposed model, the conjugate priors have provided more robust modeling for the sparse data set. For further studies a quantization or thresholding of the rating scale is proposed, e.g. the scale is mapped from $\{1,2,3,4,5\} \rightarrow \{-1,1\}$. Thus, the hope is to counteract the fact that users utilizes the scale differently by modeling in general higher ratings as likes and lower ratings as dislikes. Moreover to make the model evaluation procedure more intuitive, a recommendation metric that predicts user-personalized ranking lists should be used instead of measuring the prediction errors.

5. REFERENCES

[1] N. Barbierir, G. Manco, and E. Ritacco, *Probabilistic Approaches to Recommendations*. Morgan & Claypool, 2014.

[2] T. Hofmann, "Latent Semantic Models for Collaborative Filtering," *ACM Trans. Inf. Sys.*, vol. 22, pp. 89–115, Jan. 2004.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] J. Chu and Y. Lee, "Conjugate Prior Penalized Learning of Gaussian Mixture Models for EMG Pattern Recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (San Diego, CA), pp. 1093–1098, 2007.

[5] J. Chu and Y. Lee, "Conjugate-Prior-Penalized Learning of Gaussian Mixture Models for Multifunction Myoelectric Hand Control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, pp. 287–297, June 2009.

[6] J. Chien and M. Wu, "Adaptive Bayesian Latent Semantic Analysis," vol. 16, pp. 198–207, January 2008.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds., *Recommender Systems Handbook*. Springer, 2011.

[9] D. Ormoneit and V. Tresp, "Averaging, Maximum Penalized Likelihood and Bayesian Estimation for Improving Gaussian Mixture Probability Density Estimates," vol. 9, pp. 639–650, July 1998.

[10] D. E. Corporation, "EachMovie recommendation data set." "`http://www.gatsby.ucl.ac.uk/~chuwei/data/EachMovie/eachmovie.html`", 2004.