Factorization Techniques for Student Performance Classification and Ranking

Lucas Drumond, Nguyen Thai-Nghe, Tomáš Horváth, Lars Schmidt-Thieme

Information Systems and Machine Learning Lab,
University of Hildesheim, Germany
{ldrumond,nguyen,horvath,schnmidt-thieme}@ismll.de
www.ismll.de

Abstract. Historically, student performance prediction has been approached with regression models. For instance, the KDD Cup 2010 used the root mean squared error (RMSE) as an evaluation criterion. This is appropriate when the goal is to predict student marks or how well will they perform in a given exercise. Since in many datasets the target variable is binary, i.e. a student has solved the exercise or failed in it, it would be natural to look at this problem as to a classification task. Another, probably not so usual case could be when we only have a so-called positive feedback, i.e. only the successful solutions are recorded. In this case, neither the regression nor the classification approaches would be useful and one could look on this problem as to a ranking task. We propose to look at solving the student performance prediction as a classification or ranking tasks, respectively, where models are optimized for appropriate error measures which are the Hinge loss and the Area under the ROC curve. Experimental comparison of these techniques are introduced using two, large-scale datasets. Both methods are well known in their respective fields, thus the goal of this paper is to introduce them in the educational data mining community.

Key words: Matrix Factorization, Student Modeling, Personalization, Bayesian Personalized Ranking

1 Introduction

To address the student performance prediction problem, many works have been published so far. Most of them relying on traditional methods such as logistic regression [1], linear regression [2], decision trees [3], neural networks [4], support vector machines [5], hidden Markov models [6], Bayesian networks [7], and so on.

Matrix factorization via Stochastic Gradient Descent algorithm [8] is one of the most prominent approaches among the recommender system techniques, applied successfully also in educational data mining [9, 10]. An important attribute of matrix factorization techniques is that these result in very precise models even if no additional attributes (meta-data of students and tasks) are known and the models are learned only from the (student-task-outcome) triples.

Usually, as was also in the KDD Cup 2010, models for student performance prediction are evaluated using the root mean squared error (RMSE) suggesting that it is a regression task. This is straightforward, especially when the goal is to predict student marks or how well will they perform in a given exercise.

Inspired by the dataset from the KDD Cup in 2010 where the target labels were binary, i.e. a student was or was not successful in solving the exercise or the particular steps of an exercise, we can look to student performance prediction also as to a classification task. Moreover, it can also happen, that some system records only the successful solutions. Probably, it is not so usual but still remains as an option. In such case, we have only one target label what limits the use of regression or classification techniques. This situation is equivalent to the case of learning from positive-only feedback – a well known concept in the recommender system community. We can model this situation in the context of ranking.

The goal of this paper is to propose a different perspective to the student performance prediction problem as opposed to the regression formulation that is common in the literature. First this problem can be viewed as a classification task. We show how the problem can be approached from this perspective by optimizing a matrix factorization model for a smooth version of the Hinge Loss [11], a loss function widely used for learning classification models. Student performance prediction can also be seen as ranking problem. To couple with this view, we optimize a factorization model for the Bayesian Personalized Ranking (BPR) optimization criterion [12], a smooth approximation to the Area Under the ROC Curve (AUC), that has been successfully applied for the item prediction task in recommender systems.

2 Matrix Factorization Models

Student performance prediction task is to predict whether a given student will correctly solve a given problem, given his/her performance on other problems. The performance of a student in a problem is usually modeled as a binary variable $y_{sp} \in \{-1.+1\}$. $y_{sp} = +1$ indicates that the student s correctly solved problem p. As already shown in the literature, this problem can be successfully approached using factorization models [9]. A factorization model maps each student s and problem p to a respective k-dimensional latent feature vector, i.e. $\mathbf{s} \in \mathbb{R}^k$ and $\mathbf{p} \in \mathbb{R}^k$. The actual performance of a student is predicted as

$$\hat{y}_{sp} := \mathbf{s}\mathbf{p}^\top$$

The latent feature vectors are learned by optimizing a regularized loss function that can be generally written as:

$$Loss := \mathcal{L}(\mathbf{Y}, \mathbf{S}\mathbf{P}^\top) + \lambda(||S||^2 + ||P||^2)$$

where $\mathbf{S} \in \mathbb{R}^{|S| \times k}$ and $\mathbf{P} \in \mathbb{R}^{|P| \times k}$ are matrices where each row is the latent feature vector of a given student and problem respectively; $\mathbf{Y}$ is a matrix containing the actual performance of the different students and λ is a regularization term (hyper-parameter) to prevent the over-fitting of the model. Since $\mathbf{Y}$ is usually partially observed, $\mathcal{L}$ is a function that evaluates how good the factorization model approximates the observed part of $\mathbf{Y}$. There are a number of possible choices for $\mathcal{L}$. In this section we discuss some possibilities not yet explored in the student performance prediction literature.

2.1 Learning Student Performance prediction with the Hinge Loss

In this section, we introduce the student performance prediction as a classification task. It makes sense to optimize the model to a loss suited for this task. One successful loss used in this area is the hinge loss, which is, for instance, the loss Support Vector Machines optimize for. The hinge loss is defined as

$$hinge(y, \hat{y}) := max(0, 1 - y\hat{y})$$

where $\hat{y}$ is the score predicted by the model. One drawback of using the Hinge Loss is that it is not smooth, thus not being easy to optimize by gradient based methods. So we use the smooth hinge loss proposed by [11] defined as

$$smooth\_hinge(y, \hat{y}) := \begin{cases} \frac{1}{2} - y\hat{y} & \text{if } y\hat{y} \leq 0, \\ \frac{1}{2}(1 - y\hat{y})^2 & \text{if } 0 < y\hat{y} < 1, \\ 0 & \text{if } y\hat{y} \geq 1 \end{cases} \tag{1}$$

2.2 Learning Student Performance prediction with the Bayesian Personalized Ranking Framework

If we are searching for a ranking model, it makes sense to learn a model able to rank the problems according to the likelihood that the student will succeed in solving it. The Bayesian Personalized Ranking [12] is a framework for learn to rank that has been successfully applied to recommender systems. It can be seen as a direct optimization of the Area Under the ROC Curve (AUC) measure. It is defined as:

$$BPR - Opt := \sum_{(s,i,j) \in D} \ln \sigma(x_{sij})$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function and $x_{sij} = \hat{y}_{si} - \hat{y}_{sj}$ with $\hat{y}_{si}, \hat{y}_{sj}$ being the predicted ranks for the items i and j, respectively, by the student s.

One crucial step in using BPR is delineating how the training data D look like. Here, a triple $(s, i, j) \in D$ refers to exercises i, j and a student s, such that the exercise i received higher rank[1] than the exercise j by the student s.

[1] In other words, the exercise i is rated/classified higher than the exercise j.

In [12], the authors show how to apply BPR to a recommendation problem where only positive observations are available. In the student performance prediction setting, however, one usually has both positive and negative observations available. Here we propose three different ways of defining the set D in order to take into account this information.

For convenience, we define B_s^+ as being the set of problems solved correctly by the student s and B_s^-, the set of problems for which the student s has attempted to solve but failed to provide a correct answer (negative observations). Finally we define the set $B_s^?$ as the problems that the student never attempted to solve. We devise three alternative definitions for D:

- $D_+ := \{(s,i,j)|i \in B_s^+ \wedge j \notin B_s^+\}$ - in line with the definition in [12], this creates a contrast between problems correctly solved by the students and the rest. While this definition makes sense for positive only feedback data, it is not a good fit for the kind of educational data we work with here, as it does not account for the negative observations.
- $D_{+/-} := \{(s,i,j)|i \in B_s^+ \wedge j \in B_s^-\}$ - Creates a contrast between problems where the students either succeeded or failed to correctly solve.
- $D_{+/-/?} := \{(s,i,j)|(i \in B_s^+ \wedge j \in B_s^- \bigcup B_s^?) \vee (i \in B_s^? \wedge j \in B_s^-)\}$ - Same as the previous contrast but also taking into account the problems for which the user never attempted to solve.

3 Experiments

In the experiments, we have used two large data sets from the KDD Challenge 2010[2]. These data, namely Algebra and Bridge, represent the log files of interactions between students and the tutoring system.

Despite there is a lot of information recorded in the dataset, we consider here only the sparse *student-task matrix* consisting of the correct first attempt (CFA) information, i.e. if a student has solved the given task[3] successfully (CFA=1) or not (CFA=0). The main characteristics of the two datasets, as the number of students, tasks and student-task-CFA triples, are presented in the table 1.

Table 1: Data set characteristics

Data set	**# Students**	**# Tasks**	**# (student,task,CFA)**
Algebra 2008-2009	3,310	2,979	8,918,054
Bridge to Algebra 2008-2009	6,043	1,458	20,012,498

Fig. 1 shows the results of optimizing a factorization model for the different criteria proposed in this paper on both datasets. For comparison, we provide

[2] http://pslcdatashop.web.cmu.edu/KDDCup/

[3] A task represents one step (part) of a problem, here.

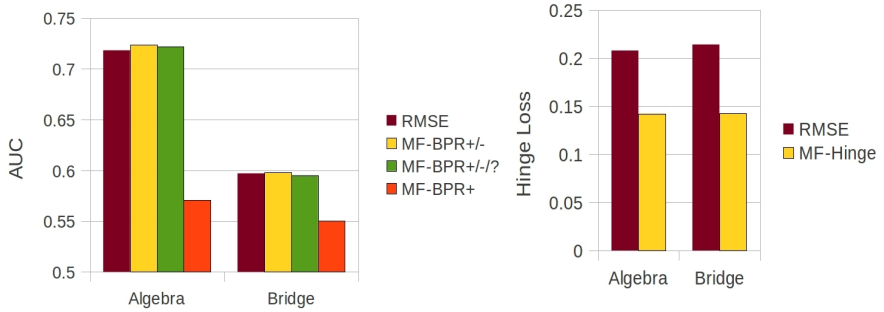
the results of optimizing a model for the root mean squared error, too. For fair comparison, we evaluated the models using a classification and a ranking measure. For classification we used Hinge loss:

$$\text{hinge}(\mathbf{y}, \hat{\mathbf{y}}) := \frac{1}{|T|} \sum_{(s,t,y_{st}) \in T} max(0, 1 - (y_{st}\hat{y}_{st}))$$

whereas for ranking we considered AUC:

$$\text{AUC}(\hat{\mathbf{y}}) := \frac{1}{|T|} \sum_{s \in T} \frac{1}{|B_s^+||B_s^-|} \sum_{i \in B_s^+} \sum_{j \in B_s^-} \delta(\hat{y}_{si} > \hat{y}_{sj})$$

where T refers to test data and $\delta(x)$ is 1 if x is true and 0 otherwise.



AUC performance of the tested ranking models. The *higher* the better.

Hinge Loss performance of the tested classification models. The *lower* the better.

Fig. 1: Results for the Hinge Loss and AUC

The reason of the poor performance of MF-BPR^+ is that when optimizing BPR one should take the negative observations into account, differentiating it from the unobserved ones. On the other hand, MF-$BPR^{+/-/?}$ performance is really close to MF-$BPR^{+/-}$ providing empirical evidence that the model does not benefit from the contrast with the unobserved cases once both positive and negative examples are available. Finally, one can see that the Hinge loss performs better than RMSE.

4 Conclusions

Two factorization techniques and their performance on large-scale datasets were presented in this paper to approach the student performance prediction problem as a classification and ranking tasks, respectively. As far as we know, these techniques were not yet introduced to the educational data mining community.

Acknowledgments: Nguyen Thai-Nghe was funded by the TRIG project of Cantho university, Vietnam. Lucas Drumond was funded by the CNPq, Brazil. Tomáš Horváth was also supported by the grant VEGA 1/0832/12 of the Pavol Jozef Šafárik University in Košice, Slovakia.

References

1. Cen, H., Koedinger, K., Junker, B.: Learning factors analysis and a general method for cognitive model evaluation and improvement. In: Proceeding of the International Conference on Intelligent Tutoring Systems, Springer (2006) 164–175
2. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Modeling and User-Adapted Interaction **19**(3) (2009) 243–266
3. Thai-Nghe, N., Janecek, P., Haddawy, P.: A comparative analysis of techniques for predicting academic performance. In: Proceeding of the 37th IEEE Frontiers in Education Conference, Milwaukee, USA, IEEE Xplore (2007) T2G7–T2G12
4. Romero, C., Ventura, S., Espejo, P.G., Herves, C.: Data mining algorithms to classify students. In: The 1st International Conference on Educational Data Mining, Montreal, Canada (2008) 8–17
5. Thai-Nghe, N., Busche, A., Schmidt-Thieme, L.: Improving academic performance prediction by dealing with class imbalance. In: Proceeding of the 9th IEEE International Conference on Intelligent Systems Design and Applications, Pisa, Italy, IEEE Computer Society (2009) 878–883
6. Pardos, Z.A., Heffernan, N.T.: Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. In: KDD Cup 2010 Workshop, Washington, DC, USA (2010)
7. Bekele, R., Menzel, W.: A bayesian approach to predict performance of a student (bapps): A case with ethiopian students. In: Proceedings of the International Conference on Artificial Intelligence and Applications, Vienna, Austria (2005) 189–194
8. Koren, Y.: Factor in the neighbors: Scalable and accurate collaborative filtering. ACM Transactions on Knowledge Discovery from Data **4**(1) (2010)
9. Thai-Nghe, N., Drumond, L., Horváth, T., Krohn-Grimberghe, A., Nanopoulos, A., Schmidt-Thieme, L.: Factorization techniques for predicting student performance. In: Educational Recommender Systems and Technologies: Practices and Challenges. IGI Global (2011)
10. Töscher, A., Jahrer, M.: Collaborative filtering applied to educational data mining. In: Proceedings of the KDD Cup 2010 Workshop, Washington, DC, USA (2010)
11. Rennie, J.: Smooth hinge classification (February 2005)
12. Rendle, S., Freudenthaler, C., Gantner, Z., Lars, S.T.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, AUAI Press (2009)