Using Recurrent Neural Networks to Build a Stopping Algorithm for an Adaptive Assessment

Jeffrey Matayoshi[1⋆], Eric Cosyn[1], and Hasan Uzun[1]

McGraw-Hill Education/ALEKS Corporation, Irvine, CA, USA
{jeffrey.matayoshi,eric.cosyn,hasan.uzun}@aleks.com

Abstract. ALEKS ("**A**ssessment and **LE**arning in **K**nowledge **S**paces") is an adaptive learning and assessment system based on knowledge space theory. In this work, our goal is to improve the overall efficiency of the ALEKS assessment by developing an algorithm that can accurately predict when the assessment should be stopped. Using data from more than 1.4 million assessments, we first build recurrent neural network classifiers that attempt to predict the final result of each assessment. We then use these classifiers to develop our stopping algorithm, with the test results indicating that the length of the assessment can potentially be reduced by a large amount while maintaining a high level of accuracy.

Keywords: Recurrent neural networks · Adaptive assessment · Knowledge space theory · Deep learning

1 Introduction and Background

ALEKS ("**A**ssessment and **LE**arning in **K**nowledge **S**paces") is a web-based, artificially intelligent system [17] based on knowledge space theory (KST) [5–7]. The foundation of ALEKS is an adaptive assessment that aims to precisely and efficiently identify the topics in an academic course that a student knows. ALEKS Placement, Preparation and Learning (ALEKS PPL) is a specialized product that has been developed to offer recommendations for placing students in post-secondary mathematics courses.

Deep learning has recently achieved dramatic successes in various fields [14] and is beginning to move into the education domain. In particular, because of the sequential nature of many types of educational data, recurrent neural networks (RNNs) are appearing more frequently in the educational literature [1, 11–13, 15, 18, 21]. Our goal is to augment the performance and efficiency of the KST-powered adaptive assessment algorithm of ALEKS PPL with the classification strengths of RNN models.

In KST, an *item* is a problem type that tests a discrete unit of the curriculum. A *knowledge state* is a set of items that a student masters, and a *knowledge space* is the collection of all such feasible knowledge states. At all times in an ALEKS PPL assessment, the 314 items under consideration are partitioned into the following categories:

⋆ Corresponding author.

– items that are most likely in the student's knowledge state (in-state);
– items that are most likely not in the student's knowledge state (out-of-state);
– the remaining items (uncertain).

The assessment terminates when either (a) there are no remaining "uncertain" items, or (b) the predetermined limit of 29 questions is reached.[1] The assessment then returns the in-state items as its best estimate of the student's knowledge state. Most ALEKS PPL assessments reach the maximum limit of 29 questions and thus end with a number of "uncertain" items. The *percentage score* of the student is simply the percentage of the 314 items that are categorized as being in-state. Based on the value of the percentage score at the end of the assessment, ALEKS PPL recommends placement in one of six different mathematics courses (see [4] for further details and background on ALEKS PPL).

2 Experimental Setup and Models

The data for our experiments consist of 1,449,625 full-length (i.e., 29 question) ALEKS PPL assessments, with each assessment being taken by a unique student for placement purposes in a college or university setting. We use 50,000 assessments for a held-out test set, another 50,000 for a validation set to tune hyperparameters and compare several models, and the remainder (1,349,625) for training our models. Each assessment generates a sequence of inputs, and the target (ground truth) label for each sequence is determined by the course placement recommendation made by the ALEKS system using all 29 questions from the assessment. Thus, the results of the ALEKS PPL assessment can be viewed as a multiclass classification problem with six different class labels, one for each of the possible course placement recommendations.

For our RNN models, we use two different recurrent units: gated recurrent units (GRU) [2] and long short-term memory (LSTM) units [9]. We include both models in our experiments since there currently is not a consensus that one architecture or the other gives superior performance, as several studies have not revealed a clear winner; these include studies both within the education domain [1, 12], as well as from the broader AI community [3, 22]. Additionally, as a comparison, we also build a set of logistic regression classifiers.

Our models will use the actual item categorizations of the ALEKS assessment as features. Thus, we require $3 \times 314 = 942$ independent variables to represent all possible combinations of assessment categories (in-state, out-of-state, and uncertain) and items. The n-th vector of each sequence contains the categorization of the items by the assessment after question n.

For the LSTM and GRU models, the number of hidden layers, the sizes of the hidden layers, and the learning rate are tuned on the validation set. We also use

[1] Students actually answer up to 30 questions when accounting for a randomly chosen question that is used for validation and other statistics. This number of questions balances the need to gather enough information about the student's knowledge state against the possibility of overwhelming the student. Regarding the latter concern, see [16] for evidence of a "fatigue effect" experienced by students in ALEKS assessments.

batch normalization [10] and, to help prevent overfitting, early stopping [19] and dropout [8, 20]. For the logistic regression models, the only tuned hyperparameter is the strength of the L2 regularization.

3 Stopping Algorithm and Model Evaluation

The best performing models on the validation set are used to implement our stopping algorithm for the ALEKS assessment. As shown in Algorithm 1, our first criterion is that the most confident predicted class label is above a certain threshold, α. Additionally, we require that the course placement recommendation at the current question (as determined by the student's percentage score at that point in the assessment) matches the classifier's predicted class label, and we also require that the assessment has asked at least 10 questions (to ensure that our classifier has a minimal amount of data to work with).

Algorithm 1 Assessment stopping algorithm

Inputs:
α, stopping threshold probability
$P(k \mid \mathbf{x}_n)$, predicted probability of class k, $k = 1, \ldots, 6$, after question n
$K_n = \arg\max_{k=1,\ldots,6} P(k \mid \mathbf{x}_n)$; i.e., the most likely class after question n
C_n, the current recommended course placement after question n

Iterations:
for $n = 10$ to 29 **do**
  Compute K_n and C_n using information from questions 1 to n
  if n == 29 or $\big(P(K_n \mid \mathbf{x}_n) > \alpha$ and $K_n == C_n\big)$ **then**
    Stop the assessment
  end if
end for

Output:
C_n, the (predicted) course placement recommendation

The results from applying Algorithm 1 to the held-out test data are shown in Figure 1, where we plot the average assessment length versus the accuracy of the predicted course placement recommendation, for various probability thresholds (i.e., various values of α). The plot shows that at any accuracy rate of 0.995 or higher, the RNN models are a minimum of 1.5 questions better than the logistic regression, with the maximum difference being about 2.2 questions.

Next, Table 1 shows the results for the LSTM RNN model partitioned by the actual (ground truth) classification label, using a value of $\alpha = 0.99$. The best results are for the extreme labels 1 and 6; on the other hand, while still being acceptable, the gains are not nearly as large for labels 4 and 5. It is worth mentioning that these results closely parallel what was found in [4], where it was

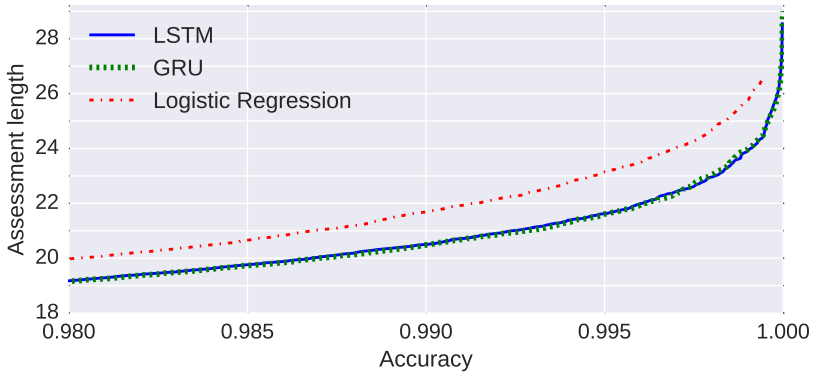

Fig. 1. Average assessment length vs. accuracy on held-out test data.

shown that ALEKS PPL has the greatest variability for labels 4 and 5, and it seems likely that this variability is a major reason for the weaker performance of the stopping algorithm with these labels.

Table 1. Stopping statistics by ground truth label for the LSTM RNN model on held-out test data, using a threshold of $\alpha = 0.99$.

Class label	1	2	3	4	5	6
Sample size	4357	8680	11108	7640	8259	9956
Average length	17.87	21.74	22.25	24.75	25.8	16.54
Accuracy	0.9963	0.9955	0.9959	0.9921	0.9921	0.9971

4 Discussion

The results from applying our stopping algorithm on a held-out test set show a large potential reduction in the average length of the ALEKS PPL assessment. For example, Figure 1 shows that at an accuracy of 0.995 the average number of questions for the RNN models is about 21.6, a roughly 25% reduction from the full-length assessment of 29 questions. Additionally, the GRU and LSTM models perform equally well, with both outperforming the logistic regression model, adding further evidence to the growing literature supporting the benefits of applying RNN models to educational data. Of note is that we use a relatively general approach, in that the features are obtained simply by taking the output of the assessment and feeding it to an RNN. The effectiveness of this technique here motivates the need for further studies involving other adaptive assessments; at the moment, it is not clear if this approach can be successful more generally, or if it is some peculiarity of ALEKS PPL that allows it to work so well.

References

1. Botelho, A., Baker, R., Heffernan, N.: Improving sensor-free affect detection using deep learning. In: Artificial Intelligence in Education-18th International Conference, AIED 2017. pp. 40–51 (2017)
2. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR **abs/1406.1078** (2014), http://arxiv.org/abs/1406.1078
3. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
4. Doble, C., Matayoshi, J., Cosyn, E., Uzun, H., Karami, A.: A data-based simulation study of reliability for an adaptive assessment based on knowledge space theory. International Journal of Artificial Intelligence in Education (2019). https://doi.org/10.1007/s40593-019-00176-0
5. Doignon, J.P., Falmagne, J.C.: Spaces for the assessment of knowledge. International Journal of Man-Machine Studies **23**, 175–196 (1985)
6. Falmagne, J.C., Albert, D., Doble, C., Eppstein, D., Hu, X. (eds.): Knowledge Spaces: Applications in Education. Springer-Verlag, Heidelberg (2013)
7. Falmagne, J.C., Doignon, J.P.: Learning Spaces. Springer-Verlag, Heidelberg (2011)
8. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. In: Advances in Neural Information Processing Systems 29 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**, 1735–1780 (1997)
10. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015)
11. Jiang, W., Pardos, Z.A., Wei, Q.: Goal-based course recommendation. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. pp. 36–45 (2019)
12. Jiang, Y., Bosch, N., Baker, R.S., Paquette, L., Ocumpaugh, J., Andres, J.M.A.L., Moore, A.L., Biswas, G.: Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? In: Artificial Intelligence in Education-19th International Conference, AIED 2018. pp. 198–211 (2018)
13. Khajah, M., Lindsey, R., Mozer, M.: How deep is knowledge tracing? In: Proceedings of the 9th International Conference on Educational Data Mining. pp. 94–101 (2016)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
15. Lin., C., Chi, M.: A comparison of BKT, RNN and LSTM for learning gain prediction. In: Artificial Intelligence in Education-18th International Conference, AIED 2017. pp. 536–539 (2017)
16. Matayoshi, J., Granziol, U., Doble, C., Uzun, H., Cosyn, E.: Forgetting curves and testing effect in an adaptive learning and assessment system. In: Proceedings of the 11th International Conference on Educational Data Mining. pp. 607–612 (2018)
17. McGraw-Hill Education/ALEKS Corporation: What is ALEKS? https://www.aleks.com/about_aleks
18. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., Sohl-Dickstein, J.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems. pp. 505–513 (2015)

19. Prechelt, L.: Early stopping – but when? In: Montavon, G., Orr, G., Müller, K. (eds.) Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol. 7700. Springer, Berlin, Heidelberg (2012)
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1968 (2014)
21. Xiong, X., Zhao, S., Vaninwegen, E., Beck, J.: Going deeper with knowledge tracing. In: Proceedings of the 9th International Conference on Educational Data Mining. pp. 545–550 (2016)
22. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923 (2017)