Machine Learning Engineer Nanodegree

Capstone Proposal

Determination of students' interaction patterns with an intelligent tutoring system and study of their correlation with successful learning

Dmytro Iakubovskyi, October 6, 2017

Domain Background

How to make the learning process more successful? As a passionate learner, I am interested to find out the answer. Usually, the process of learning is described by the learning curve. Numerous studies of different types of learning suggest the *scale-invariant* learning curve defined by the "Power Law of Learning":

$$E(x) = b\, x^{-d} \,,$$

where $\boldsymbol{x}$ is the number of times student received feedback, $\boldsymbol{E(x)}$ is the average error rate, $\boldsymbol{b}$ is the *difficulty parameter* and $\boldsymbol{d}$ is the *learning parameter*. Despite its simplicity, the Power Law of Learning describes the observed variety of learning curves remarkably well, see [1-6].

In this project, I will study the learning curves provided by an intelligent tutoring system (ITS). ITS are designed to give the learners an instantaneous automatic feedback concerning the correctness of their solutions, see, e.g., [7] for an ITS review. Many previous works (see, e.g., [8-10] for reviews) have used machine learning to analyse the ITS output.

Problem Statement

Students with *higher* learning parameter should have *more favourabl*e learning. I hypothesise that the students *often* communicate with the tutor (e.g., in the form of an intelligent tutoring system) not in an optimal way. Therefore, the problem can be split into *three steps*:

1. to *cluster* students based on their study patterns;
2. to *determine* the value of learning parameter for each student;
3. to *rank* the obtained clusters based on learning parameter distributions within them.

Datasets and Inputs

Modern ITS can provide large amounts of relevant data. In this project, I am going to investigate the *ASSISTments* dataset [11] publicly available at PSLC DataShop [12]. It contains detailed records from several hundreds of schoolchildren regularly interacting with an intelligent tutoring system (in addition to regular classes) during 2004-2005, 2005-2006 and 2006-2007 school years. This dataset is large enough for a detailed statistical analysis. In total, it contains about 3.3M instances collected from several thousands of students, and 100+ features (including all information necessary to determine our target variable — the learning parameter). The critical element of the *ASSISTments* system is that it proposes several "scaffolding" questions to coach students that failed on original questions.

The *ASSISTments* dataset has been already used in some related studies. For example, [13, 14] reported that despite very limited interaction with the system (one 20-minute or 40-minute lesson every two weeks [15], depending on the school), the students improved their knowledge during such interaction. Also, [15-17] found that the models based on *skill learning tracking* predicted the final test score better than a simple model based only on time variable and the difficulty parameter $\boldsymbol{b}$. Also, [18] studied in details the "gaming" phenomenon where "a learner attempts to succeed in an educational environment by exploiting properties of the system's help and feedback rather than by attempting to learn the material".

Solution Statement

Step 1 is a typical clustering (unsupervised classification) problem. For this action, I will use the variety of clustering algorithms implemented in the scikit-learn library for machine learning and described in details here. Due to limited complexity of the problem (< 10 000 samples and an unknown number of desired clusters), MeanShift or VBGMM algorithms are expected to be the best. However, I am going to compare their performance with other widely used clustering algorithms (such as KMeans) and to choose the algorithm with the largest Silhouette Coefficient selected as a relevant evaluation metrics, see below.

Step 2 requires determination of the learning parameters for participants within each cluster. Because of the Power Law of Learning, we expect that the error rate ***E(x)*** for each student will be the *power law* function of the number of times ***x*** student received feedback. The log-log transformation is feasible because after doing that, the problem will be reformulated as a simple Ordinary Least Squares regression. To robustly estimate the coefficient values, I will also try alternative algorithms, such as Ridge Regression and Lasso, and choose the algorithm with the largest R2 score selected as a relevant evaluation metrics, see below.

Finally, step 3 does not need machine learning algorithms only using the outcomes of steps 1 and 2.

Benchmark Model

I select the benchmark model based on the finding of [18] that some students showed explicit "gaming" behaviour when using ASSISTments. An example of "gaming" action is the student Mary who, according to [13], "used the system 4 hours and 17 minutes, finished 114 items with 20% correct ... went through 356 scaffolding questions with 20% correct and asked for 705 hints, which is enormous compared to her classmates". As a result, students could be grouped into at least *two* clusters - “gaming” vs “non-gaming” students. I will try to check this hypothesis by running KMeans algorithm with only two clusters. According to [18], I expect that one of these clusters will correspond to "gaming" behaviour (i.e., fast and low-performance interactions with the system).

Evaluation Metrics

For step 1, the clusters will be selected to maximise the Silhouette Coefficient which is the average value of individual scores ***s*** calculated for each sample:

$$s = \frac{b - a}{max(a, b)}$$

where ***a*** is the mean distance between a sample and all other points in the same class, ***b*** is the average distance between a sample and all other points in the *next nearest cluster*. This score is *bounded* from -1 and +1 and is higher for dense and well-separated clusters, similar to expectations.

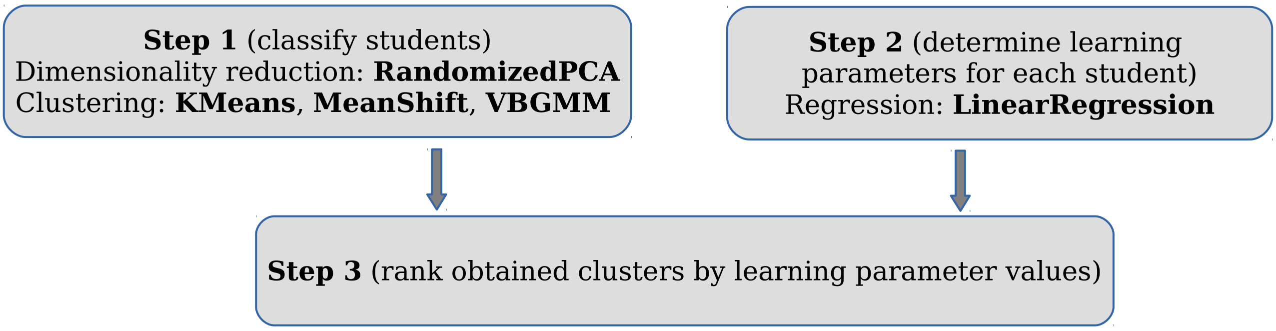
For step 2, the LinearRegression, Ridge and Lasso algorithms will be compared, and the best algorithm selection will be based on the usual R2 score:

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}.$$

Here, SSres is the residual sum of squares, SStot is the total sum of squares.

Project Design

The project workflow can be summarised as follows:



As seen from the above Figure, the project consists of three stages. The first two steps are independent, and the third is a compilation of them.

An outcome of step 1 is the list of clusters grouped by student interaction patterns with the *ASSISTments* system. Because clustering algorithms usually work well in a small number of categories, the pre-processing stage reducing the dimensionality of the dataset is required. As described in Solution Statement Section, I will compare the performance KMeans, MeanShift and VBGMM clustering algorithms, and will choose the one with the largest Silhouette Coefficient. Also, before clustering the data, the suitable dimensionality reduction is required. According to the scikit-learn algorithm cheat-sheet, RandomizedPCA algorithm seems the most appropriate.

An outcome if step 2 is the values of learning parameters for each student. After log-log transformation, it will be reformulated as a Linear Regression problem and solved with LinearRegression, Ridge and Lasso algorithms, the best of them will be chosen given the largest R2 score.

Finally, as a result of *step 3*, the clusters obtained during *step 1* will be graded according to the learning parameter values calculated during *step 2*. Both mean value and standard deviation of the leaning parameter will be used to answer whether a given cluster describes better learning compared to others.

References:

[1] Snoddy, G. S. (1926). *Learning and stability: a psychophysiological analysis of a case of motor learning with clinical applications*. Journal of Applied Psychology, 10(1):1 – 36.

[2] Crossman, E. R. F. W. (1959). *A theory of the acquisition of speed-skill*. Ergonomics, 2(2):153–166.

[3] Card, S. K., Moran, T. P., and Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates, Hillsdale, NJ.

[4] Fitts, P. M. and Posner, M. I. (1967). *Human performance*. Brooks Cole, Belmont, CA.

[5] Newell, A. and Rosenbloom, P. S. (1981). *Mechanisms of skill acquisition and the law of practice*. In Anderson, J. R., editor, Cognitive skills and their acquisition, pages 1–55. Lawrence Erlbaum Associates, Hillsdale, NJ.

[6] Anderson, J. R. (1982). *Acquisition of cognitive skill*. Psychological Review, 89(4):369–406.

[7] Nwana, H. S. (1990). *Intelligent Tutoring Systems: an overview*. Artificial Intelligence Review, 4, 251-277.

[8] Hämäläinen, W., Vinni, M. (2006) *Comparison of Machine Learning Methods for Intelligent Tutoring Systems*. In: Ikeda M., Ashley K.D., Chan TW. (eds) Intelligent Tutoring Systems. ITS 2006. Lecture Notes in Computer Science, vol 4053. Springer, Berlin, Heidelberg.

[9] Romero, C., & Ventura, S. (2010). *Educational Data Mining: A Review of the State of the Art*. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 40. 601 - 618.

[10] Shahiri, A. M., Wahidah, H., Nur'aini, A.R. *A Review on Predicting Student's Performance Using Data Mining Techniques*. Procedia Computer Science, Volume 72, 2015, Pages 414-422.

[11] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., & Rasmussen, K.P. (2005). *The Assistment Project: Blending Assessment and Assisting*. In C.K. Looi, G.McCalla, B. Bredeweg, & J. Breuker (Eds.) Proceedings of the 12th Artificial Intelligence in Education. Amsterdam: ISO Press. pp. 555–562.

[12] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010) *A Data Repository for the EDM community*: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

[13] Feng, M., Heffernan, N.T, Koedinger, K.R. (2006). *Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses*. In Proceedings of the 15th International World Wide Web Conference. pp. 307-316. ACM Press: New York, NY. 2006.

[14] Feng, M., Heffernan, N., Beck, J, & Koedinger, K. (2008) *Can we predict which groups of questions students will learn from?* In Baker & Beck (Eds.). Proceedings of the 1st International Conference on Education Data Mining. pp.218-225. Montreal, 2008.

[15] Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). *Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required*. In Ikeda, Ashley & Chan (Eds.) Proceedings of the Eighth International Conference on Intelligent Tutoring Systems. Springer-Verlag: Berlin. pp 31–40.

[16] Feng, M., Heffernan, N.T, Mani, M., & Heffernan C. (2006). *Using Mixed Effects Modeling to Compare Different Grain-Sized Skill Models*. In Beck, J., Aimeur, E., & Barnes, T. (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 57-66.

[17] Feng, M., Heffernan, N. (2007) *Assessing Students' Performance Longitudinally: Item Difficulty Parameter vs. Skill Learning Tracking*. Paper presented at the National Council on Educational Measurement 2007 Annual Conference, Chicago.

[18] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K. (2008) *Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments*. Journal of Interactive Learning Research, v19 n2 p185-224