

Mathematical Language Processing: Automatic Grading and Feedback for Open Response Mathematical Questions

Andrew S. Lan, Divyanshu Vats, Andrew E. Waters, Richard G. Baraniuk
Rice University
Houston, TX 77005
{mr.lan, dvats, waters, richb}@sparfa.com

ABSTRACT

While computer and communication technologies have provided effective means to scale up many aspects of education, the submission and grading of assessments such as homework assignments and tests remains a weak link. In this paper, we study the problem of automatically grading the kinds of *open response mathematical questions* that figure prominently in STEM (science, technology, engineering, and mathematics) courses. Our data-driven framework for *mathematical language processing* (MLP) leverages solution data from a large number of learners to evaluate the correctness of their solutions, assign partial-credit scores, and provide feedback to each learner on the likely locations of any errors. MLP takes inspiration from the success of natural language processing for text data and comprises three main steps. First, we convert each solution to an open response mathematical question into a series of numerical *features*. Second, we *cluster* the features from several solutions to uncover the structures of correct, partially correct, and incorrect solutions. We develop two different clustering approaches, one that leverages generic clustering algorithms and one based on Bayesian nonparametrics. Third, we *automatically grade* the remaining (potentially large number of) solutions based on their assigned cluster and one instructor-provided grade per cluster. As a bonus, we can track the cluster assignment of each step of a multistep solution and determine when it departs from a cluster of correct solutions, which enables us to indicate the *likely locations of errors* to learners. We test and validate MLP on real-world MOOC data to demonstrate how it can substantially reduce the human effort required in large-scale educational platforms.

Author Keywords

Automatic grading, Machine learning, Clustering, Bayesian nonparametrics, Assessment, Feedback, Mathematical language processing

INTRODUCTION

Large-scale educational platforms have the capability to revolutionize education by providing inexpensive, high-quality learning opportunities for millions of learners worldwide. Examples of such platforms include massive open online courses (MOOCs) [6, 7, 9, 10, 16, 42], intelligent tutoring systems [43], computer-based homework and testing systems [1, 31, 38, 40], and personalized learning systems [24]. While computer and communication technologies have provided effective means to scale up the number of learners viewing lectures (via streaming video), reading the textbook (via the web), interacting with simulations (via a graphical user interface), and engaging in discussions (via online forums), the submission and grading of assessments such as homework assignments and tests remains a weak link.

There is a pressing need to find new ways and means to automate two critical tasks that are typically handled by the instructor or course assistants in a small-scale course: (i) grading of assessments, including allotting partial credit for partially correct solutions, and (ii) providing individualized feedback to learners on the locations and types of their errors.

Substantial progress has been made on automated grading and feedback systems in several restricted domains, including essay evaluation using natural language processing (NLP) [1, 33], computer program evaluation [12, 15, 29, 32, 34], and mathematical proof verification [8, 19, 21].

In this paper, we study the problem of automatically grading the kinds of *open response mathematical questions* that figure prominently in STEM (science, technology, engineering, and mathematics) education. To the best of our knowledge, there exist no tools to automatically evaluate and allot partial-credit scores to the solutions of such questions. As a result, large-scale education platforms have resorted either to oversimplified multiple choice input and binary grading schemes (correct/incorrect), which are known to convey less information about the learners' knowledge than open response questions [17], or peer-grading schemes [25, 26], which shift the burden of grading from the course instructor to the learners.¹

¹While peer grading appears to have some pedagogical value for learners [30], each learner typically needs to grade several solutions from other learners for each question they solve, in order to obtain an accurate grade estimate.

$ \begin{aligned} & ((x^3 + \sin x)/e^x)' \\ &= ((x^3 + \sin x)e^{-x})' \\ &= (x^3 + \sin x)'e^{-x} + (x^3 + \sin x)(e^{-x})' \\ &= (3x^2 + \cos x)e^{-x} + (x^3 + \sin x)(-e^{-x}) \\ &= (3x^2 + \cos x - x^3 - \sin x)e^{-x} \\ &= \frac{3x^2 + \cos x - x^3 - \sin x}{e^x} \end{aligned} $	$ \begin{aligned} & ((x^3 + \sin x)/e^x)' \\ &= \frac{(3x^2 + \cos x)e^x - (x^3 + \sin x)e^x}{e^{2x}} \\ &= \frac{2x^2 - x^3 + \cos x - \sin x}{e^x} \end{aligned} $
(a) A correct solution that receives 3/3 credits	(b) An incorrect solution that receives 2/3 credits due to an error in the last expression
$ \begin{aligned} & ((x^3 + \sin x)/e^x)' \\ &= ((x^3 + \sin x)e^{-x})' \\ &= -e^x(x^3 + \sin x) + e^{-x}(3x^2 + \cos x) \\ &= -e^x x^3 - e^x \sin x + 3e^{-x}x^2 + e^{-x} \cos x \end{aligned} $	$ \begin{aligned} & ((x^3 + \sin x)/e^x)' \\ &= x \end{aligned} $
(c) An incorrect solution that receives 1/3 credits due to an error in the second expression	(d) An incorrect solution that receives 0/3 credits

Figure 1: Example solutions to the question “Find the derivative of $(x^3 + \sin x)/e^x$ ” that were assigned scores of 3, 2, 1 and 0 out of 3, respectively, by our MLP-B algorithm.

$ \begin{aligned} & (x^2 + x + \sin^2 x + \cos^2 x)(2x - 3) \\ &= (x^2 + x + 1)(2x - 3) \\ &= 2x^3 - 3x^2 + 2x^2 - 3x + 2x - 3 \\ &= 2x^3 - x^2 - x - 3 \end{aligned} $	$ \begin{aligned} & (x^2 + x + \sin^2 x + \cos^2 x)(2x - 3) \\ &= 2x^3 + 2x^2 + 2x \sin^2 x + 2x \cos^2 x \\ &\quad - 3x^2 - 3x - 3 \sin^2 x - 3 \cos^2 x \\ &= 2x^3 - x^2 - 3x + 2x(\sin^2 x + \cos^2 x) \\ &\quad - 3(\sin^2 x + \cos^2 x) \\ &= 2x^3 - x^2 - 3x + 2x(1) - 3(1) \\ &= 2x^3 - x^2 - x - 3 \end{aligned} $
(a) A correct solution that makes the simplification $\sin^2 x + \cos^2 x = 1$ in the first expression	(b) A correct solution that makes the simplification $\sin^2 x + \cos^2 x = 1$ in the third expression

Figure 2: Examples of two different yet correct paths to solve the question “Simplify the expression $(x^2 + x + \sin^2 x + \cos^2 x)(2x - 3)$.”

Main Contributions

In this paper, we develop a data-driven framework for *mathematical language processing* (MLP) that leverages solution data from a large number of learners to evaluate the correctness of solutions to open response mathematical questions, assign partial-credit scores, and provide feedback to each learner on the likely locations of any errors. The scope of our framework is broad and covers questions whose solution involves one or more mathematical expressions. This includes not just formal proofs but also the kinds of mathematical calculations that figure prominently in science and engineering courses. Examples of solutions to two algebra questions of various levels of correctness are given in Figures 1 and 2. In this regard, our work differs significantly from that of [8], which focuses exclusively on evaluating logical proofs.

Our MLP framework, which is inspired by the success of NLP methods for the analysis of textual solutions (e.g., essays and short answer), comprises three main steps.

First, we convert each solution to an open response mathematical question into a series of *numerical features*. In deriving these features, we make use of symbolic mathematics to transform mathematical expressions into a canonical form.

Second, we *cluster* the features from several solutions to uncover the structures of correct, partially correct, and incorrect solutions. We develop two different clustering approaches. MLP-S uses the numerical features to define a *similarity score* between pairs of solutions and then applies a generic clustering algorithm, such as spectral clustering (SC) [22] or affinity propagation (AP) [11]. We show that MLP-S is also useful for visualizing mathematical solutions. This can help instructors identify groups of learners that make similar errors so that instructors can deliver personalized remediation. MLP-B defines a *nonparametric Bayesian model* for the solutions and applies a Gibbs sampling algorithm to cluster the solutions.

Third, once a human assigns a grade to at least one solution in each cluster, we automatically *grade* the remaining (potentially large number of) solutions based on their assigned cluster. As a bonus, in MLP-B, we can track the cluster assignment of each step in a multistep solution and determine when it departs from a cluster of correct solutions, which enables us to indicate the likely locations of errors to learners.

In developing MLP, we tackle three main challenges of analyzing open response mathematical solutions. First, solutions might contain different notations that refer to the same mathematical quantity. For instance, in Figure 1, the learners use both e^{-x} and $\frac{1}{e^x}$ to refer to the same quantity. Second, some questions admit more than one path to the correct/incorrect solution. For instance, in Figure 2 we see two different yet correct solutions to the same question. It is typically infeasible for an instructor to enumerate all of these possibilities to automate the grading and feedback process. Third, numerically verifying the correctness of the solutions does not always apply to mathematical questions, especially when simplifications are required. For example, a question that asks to simplify the expression $\sin^2 x + \cos^2 x + x$ can have both $1+x$ and $\sin^2 x + \cos^2 x + x$ as numerically correct answers, since both these expressions output the same value for all values of x . However, the correct answer is $1+x$, since the question expects the learners to recognize that $\sin^2 x + \cos^2 x = 1$. Thus, methods developed to check the correctness of computer programs and formulae by specifying a range of different inputs and checking for the correct outputs, e.g., [32], cannot always be applied to accurately grade open response mathematical questions.

Related Work

Prior work has led to a number of methods for grading and providing feedback to the solutions of certain kinds of open response questions. A linear regression-based approach has been developed to grade essays using features extracted from a training corpus using Natural Language Processing (NLP) [1, 33]. Unfortunately, such a simple regression-based model does not perform well when applied to the features extracted from mathematical solutions. Several methods have been developed for automated analysis of computer programs [15, 32]. However, these methods do not apply to the solutions

to open response mathematical questions, since they lack the structure and compilability of computer programs. Several methods have also been developed to check the correctness of the logic in mathematical proofs [8, 19, 21]. However, these methods apply only to mathematical proofs involving logical operations and not the kinds of open-ended mathematical calculations that are often involved in science and engineering courses.

The idea of clustering solutions to open response questions into groups of similar solutions has been used in a number of previous endeavors: [2, 5] uses clustering to grade short, textual answers to simple questions; [23] uses clustering to visualize a large collection of computer programs; and [28] uses clustering to grade and provide feedback on computer programs. Although the high-level concept underlying these works is resonant with the MLP framework, the feature building techniques used in MLP are very different, since the structure of mathematical solutions differs significantly from short textual answers and computer programs.

This paper is organized as follows. In the next section, we develop our approach to convert open response mathematical solutions to numerical features that can be processed by machine learning algorithms. We then develop MLP-S and MLP-B and use real-world MOOC data to showcase their ability to accurately grade a large number of solutions based on the instructor’s grades for only a small number of solutions, thus substantially reducing the human effort required in large-scale educational platforms. We close with a discussion and perspectives on future research directions.

MLP FEATURE EXTRACTION

The first step in our MLP framework is to transform a collection of solutions to an open response mathematical question into a set of numerical features. In later sections, we show how the numerical features can be used to cluster and grade solutions as well as generate informative learner feedback.

A solution to an open response mathematical question will in general contain a mixture of explanatory text and core mathematical expressions. Since the correctness of a solution depends primarily on the mathematical expressions, we will ignore the text when deriving features. However, we recognize that the text is potentially very useful for automatically generating explanations for various mathematical expressions. We leave this avenue for future work.

A workhorse of NLP is the *bag-of-words* model; it has found tremendous success in text semantic analysis. This model treats a text document as a collection of words and uses the frequencies of the words as numerical features to perform tasks like topic classification and document clustering [4, 5].

A solution to an open response mathematical question consists of a series of *mathematical expressions* that are chained together by text, punctuation, or mathematical delimiters including $=$, \leq , $>$, α , \approx , etc. For example, the solution in Figure 1(b) contains the expressions $((x^3 + \sin x)/e^x)'$, $((3x^2 + \cos x)e^x - (x^3 + \sin x)e^x)/e^{2x}$, and $(2x^2 - x^3 + \cos x - \sin x)/e^x$ that are all separated by the delimiter “=”.

MLP identifies the unique mathematical expressions contained in the learners’ solutions and uses them as features, effectively extending the bag-of-words model to use mathematical expressions as features rather than words. To coin a phrase, MLP uses a novel *bag-of-expressions* model.

Once the mathematical expressions have been extracted from a solution, we parse them using SymPy, the open source Python library for symbolic mathematics [36].² SymPy has powerful capability for simplifying expressions. For example, $x^2 + x^2$ can be simplified to $2x^2$, and $e^x x^2 / e^{2x}$ can be simplified to $e^{-x} x^2$. In this way, we can identify the equivalent terms in expressions that refer to the same mathematical quantity, resulting in more accurate features. In practice for some questions, however, it might be necessary to tone down the level of SymPy’s simplification. For instance, the key to solving the question in Figure 2 is to simplify the expression using the Pythagorean identity $\sin^2 x + \cos^2 x = 1$. If SymPy is called on to perform such a simplification automatically, then it will not be possible to verify whether a learner has correctly navigated the simplification in their solution. For such problems, it is advisable to perform only arithmetic simplifications.

After extracting the expressions from the solutions, we transform the expressions into numerical features. We assume that N learners submit solutions to a particular mathematical question. Extracting the expressions from each solution using SymPy yields a total of V unique expressions across the N solutions.

We encode the solutions in a integer-valued *solution feature matrix* $\mathbf{Y} \in \mathbb{N}^{V \times N}$ whose rows correspond to different expressions and whose columns correspond to different solutions; that is, the $(i, j)^{\text{th}}$ entry of \mathbf{Y} is given by

$$Y_{i,j} = \text{times expression } i \text{ appears in solution } j.$$

Each column of \mathbf{Y} corresponds to a numerical representation of a mathematical solution. Note that we do not consider the ordering of the expressions in this model; such an extension is an interesting avenue for future work. In this paper, we indicate in \mathbf{Y} only the presence and not the frequency of an expression, i.e., $\mathbf{Y} \in \{0, 1\}^{V \times N}$ and

$$Y_{i,j} = \begin{cases} 1 & \text{if expression } i \text{ appears in solution } j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The extension to encoding frequencies is straightforward.

To illustrate how the matrix \mathbf{Y} is constructed, consider the solutions in Figure 2(a) and (b). Across both solutions, there are 7 unique expressions. Thus, \mathbf{Y} is a 7×2 matrix, with each row corresponding to a unique expression. Letting the first four rows of \mathbf{Y} correspond to the four expressions in Figure 2(a) and the remaining three rows to expressions 2–4 in Figure 2(b), we have

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}^T.$$

²In particular, we use the `parse_expr` function.

We end this section with the crucial observation that, for a wide range of mathematical questions, many expressions will be shared across learners' solutions. This is true, for instance, in Figure 2. This suggests that there are a limited number of types of solutions to a question (both correct and incorrect) and that solutions of the same type tend to be similar to each other. This leads us to the conclusion that the N solutions to a particular question can be effectively clustered into $K \ll N$ clusters. In the next two sections, we will develop MLP-S and MLP-B, two algorithms to cluster solutions according to their numerical features.

MLP-S: SIMILARITY-BASED CLUSTERING

In this section, we outline MLP-S, which clusters and then grades solutions using a solution similarity-based approach.

The MLP-S Model

We start by using the solution features in \mathbf{Y} to define a notion of similarity between pairs of solutions. Define the $N \times N$ similarity matrix \mathbf{S} containing the pairwise similarities between all solutions, with its $(i, j)^{\text{th}}$ entry the similarity between solutions i and j

$$S_{i,j} = \frac{\mathbf{y}_i^T \mathbf{y}_j}{\min\{\mathbf{y}_i^T \mathbf{y}_i, \mathbf{y}_j^T \mathbf{y}_j\}}. \quad (2)$$

The column vector \mathbf{y}_i denotes the i^{th} column of \mathbf{Y} and corresponds to learner i 's solution. Informally, $S_{i,j}$ is the number of common expressions between solution i and solution j divided by the minimum of the number of expressions in solutions i and j . A large/small value of $S_{i,j}$ corresponds to the two solutions being similar/dissimilar. For example, the similarity between the solutions in Figure 1(a) and Figure 1(b) is $1/3$ and the similarity between the solutions in Figure 2(a) and Figure 2(b) is $1/2$. \mathbf{S} is symmetric, and $0 \leq S_{i,j} \leq 1$. Equation (2) is just one of any possible solution similarity metrics. We defer the development of other metrics to future work.

Clustering Solutions in MLP-S

Having defined the similarity $S_{i,j}$ between two solutions i and j , we now cluster the N solutions into $K \ll N$ clusters such that the solutions within each cluster have high similarity score between them and solutions in different clusters have low similarity score between them.

Given the similarity matrix \mathbf{S} , we can use any of the multitude of standard clustering algorithms to cluster solutions. Two examples of clustering algorithms are *spectral clustering* (SC) [22] and *affinity propagation* (AP) [11]. The SC algorithm requires specifying the number of clusters K as an input parameter, while the AP algorithm does not.

Figure 3 illustrates how AP is able to identify clusters of similar solutions from solutions to four different mathematical questions. The figures on the top correspond to solutions to the questions in Figures 1 and 2, respectively. The bottom two figures correspond to solutions to two signal processing questions. Each node in the figure corresponds to a solution, and nodes with the same color correspond to solutions that

belong to the same cluster. For each figure, we show a sample solution from some of these clusters, with the boxed solutions corresponding to correct solutions. We can make three interesting observations from Figure 3:

- In the top left figure, we cluster a solution with the final answer $3x^2 + \cos x - (x^3 + \sin x))/e^x$ with a solution with the final answer $3x^2 + \cos x - (x^3 + \sin x))/e^x$. Although the later solution is incorrect, it contained a typographical error where $3 * x \wedge 2$ was typed as $3 \wedge x \wedge 2$. MLP-S is able to identify this typographical error, since the expression before the final solution is contained in several other correct solutions.
- In the top right figure, the correct solution requires identifying the trigonometric identity $\sin^2 x + \cos^2 x = 1$. The clustering algorithm is able to identify a subset of the learners who were not able to identify this relationship and hence could not simplify their final expression.
- MLP-S is able to identify solutions that are strongly connected to each other. Such a visualization can be extremely useful for course instructors. For example, an instructor can easily identify a group of learners who lack mastery of a certain skill that results in a common error and adjust their course plan accordingly to help these learners.

Auto-Grading via MLP-S

Having clustered all solutions into a small number K of clusters, we assign the same grade to all solutions in the same cluster. If a course instructor assigns a grade to one solution from each cluster, then MLP-S can automatically grade the remaining $N - K$ solutions. We construct the index set \mathcal{I}_S of solutions that the course instructor needs to grade as

$$\mathcal{I}_S = \left\{ \arg \max_{i \in \mathcal{C}_k} \sum_{j=1}^N S_{i,j}, k = 1, 2, \dots, K \right\},$$

where \mathcal{C}_k represents the index set of the solutions in cluster k . In words, in each cluster, we select the solution having the highest similarity to the other solutions (ties are broken randomly) to include in \mathcal{I}_S . We demonstrate the performance of auto-grading via MLP-S in the experimental results section below.

MLP-B: BAYESIAN NONPARAMETRIC CLUSTERING

In this section, we outline MLP-B, which clusters and then grades solutions using a Bayesian nonparametrics-based approach. The MLP-B model and algorithm can be interpreted as an extension of the model in [44], where a similar approach is proposed to cluster short text documents.

The MLP-B Model

Following the key observation that the N solutions can be effectively clustered into $K \ll N$ clusters, let \mathbf{z} be the $N \times 1$ cluster assignment vector, with $z_j \in \{1, \dots, K\}$ denoting the cluster assignment of the j^{th} solution with $j \in \{1, \dots, N\}$. Using this latent variable, we model the probability of the

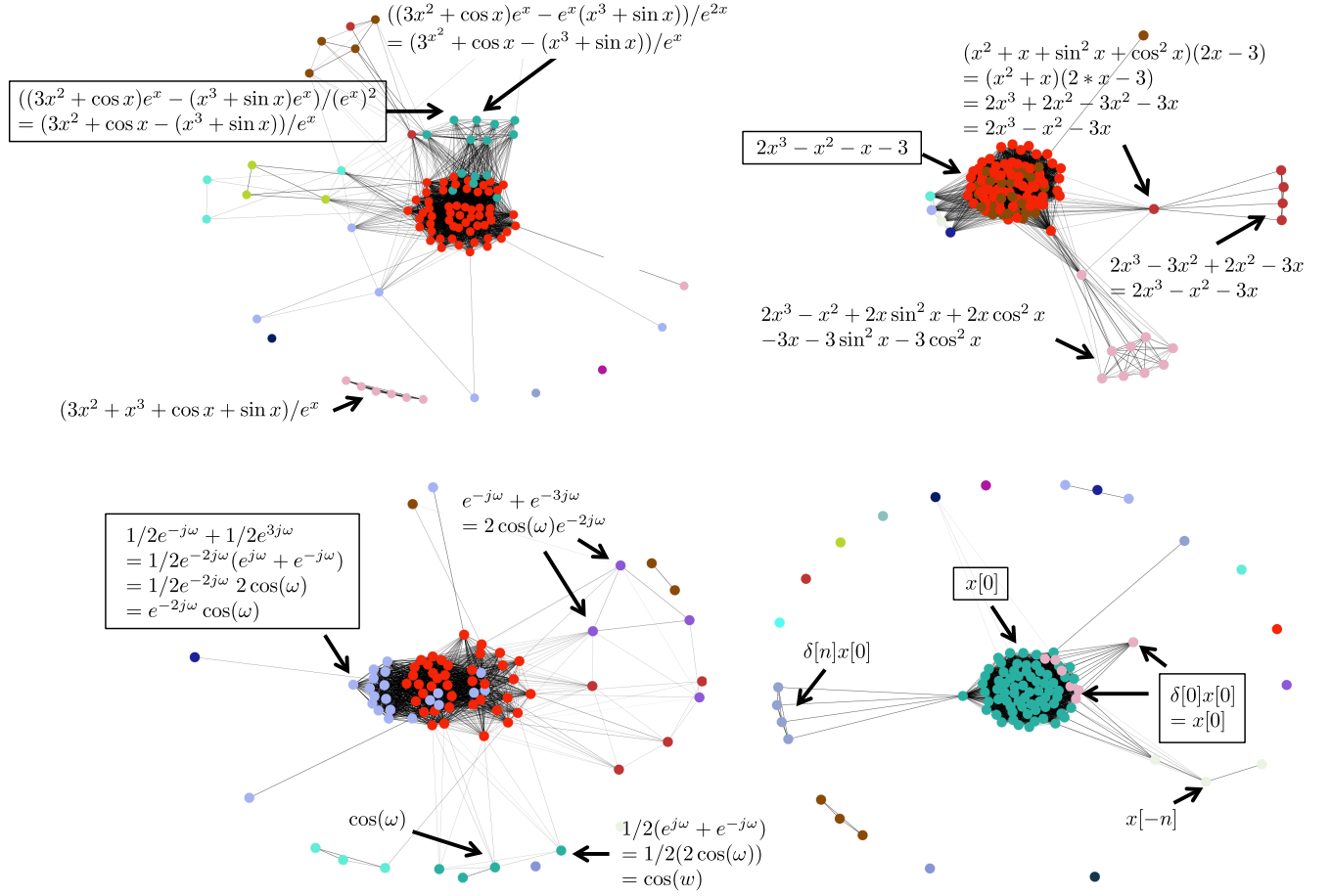


Figure 3: Illustration of the clusters obtained by MLP-S by applying affinity propagation (AP) on the similarity matrix S corresponding to learners' solutions to four different mathematical questions (see Table 1 for more details about the datasets and the Appendix for the question statements). Each node corresponds to a solution. Nodes with the same color correspond to solutions that are estimated to be in the same cluster. The thickness of the edge between two solutions is proportional to their similarity score. Boxed solutions are correct; all others are in varying degrees of correctness.

solution of all learners' solutions to the question as

$$p(\mathbf{Y}) = \prod_{j=1}^N \left(\sum_{k=1}^K p(\mathbf{y}_j | z_j = k) p(z_j = k) \right),$$

where \mathbf{y}_j , the j^{th} column of the data matrix \mathbf{Y} , corresponds to learner j 's solution to the question. Here we have implicitly assumed that the learners' solutions are independent of each other. By analogy to topic models [4, 35], we assume that learner j 's solution to the question, \mathbf{y}_j , is generated according to a *multinomial* distribution given the cluster assignments \mathbf{z} as

$$\begin{aligned} p(\mathbf{y}_j | z_j = k) &= \text{Mult}(\mathbf{y}_j | \phi_k) \\ &= \frac{(\sum_i Y_{i,j})!}{Y_{1,j}! Y_{2,j}! \dots Y_{V,j}!} \Phi_{1,k}^{Y_{1,j}} \Phi_{2,k}^{Y_{2,j}} \dots \Phi_{V,k}^{Y_{V,j}}, \end{aligned} \quad (3)$$

where $\Phi \in [0, 1]^{V \times K}$ is a parameter matrix with $\Phi_{v,k}$ denoting its $(v, k)^{\text{th}}$ entry. $\phi_k \in [0, 1]^{V \times 1}$ denotes the k^{th} column of Φ and characterizes the multinomial distribution over all the V features for cluster k .

In practice, one often has no information regarding the number of clusters K . Therefore, we consider K as an unknown parameter and infer it from the solution data. In order to do so, we impose a Chinese restaurant process (CRP) prior on the cluster assignments \mathbf{z} , parameterized by a parameter α . The CRP characterizes the random partition of data into clusters, in analogy to the seating process of customers in a Chinese restaurant. It is widely used in Bayesian mixture modeling literature [3, 14]. Under the CRP prior, the cluster (table) assignment of the j^{th} solution (customer), conditioned on the

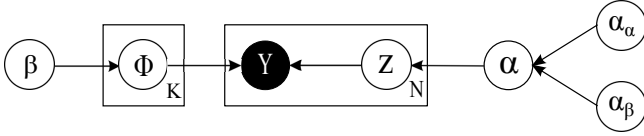


Figure 4: Graphical model of the generation process of solutions to mathematical questions. α_α , α_β and β are hyperparameters, \mathbf{z} and Φ are latent variables to be inferred, and \mathbf{Y} is the observed data defined in (1).

cluster assignments of all the other solutions, follows the distribution

$$p(z_j = k | \mathbf{z}_{-j}, \alpha) = \begin{cases} \frac{n_{k,-j}}{N-1+\alpha} & \text{if cluster } k \text{ is occupied,} \\ \frac{\alpha}{N-1+\alpha} & \text{if cluster } k \text{ is empty,} \end{cases} \quad (4)$$

where $n_{k,-j}$ represents the number of solutions that belong to cluster k excluding the current solution j , with $\sum_{k=1}^K n_{k,-j} = N - 1$. The vector \mathbf{z}_{-j} represents the cluster assignments of the other solutions. The flexibility of allowing any solution to start a new cluster of its own enables us to automatically infer K from data. It is known [37] that the expected number of clusters under the CRP prior satisfies $K \sim O(\alpha \log N) \ll N$, so our method scales well as the number of learners N grows large. We also impose a Gamma prior $\alpha \sim \text{Gam}(\alpha_\alpha, \alpha_\beta)$ on α to help us infer its value.

Since the solution feature data \mathbf{Y} is assumed to follow a multinomial distribution parameterized by Φ , we impose a symmetric Dirichlet prior over Φ as $\phi_k \sim \text{Dir}(\phi_k | \beta)$ because of its conjugacy with the multinomial distribution [13].

The graphical model representation of our model is visualized in Figure 4. Our goal next is to estimate the cluster assignments \mathbf{z} for the solution of each learner, the parameters ϕ_k of each cluster, and the number of clusters K , from the binary-valued solution feature data matrix \mathbf{Y} .

Clustering Solutions in MLP-B

We use a Gibbs sampling algorithm for posterior inference under the MLP-B model, which automatically groups solutions into clusters. We start by applying a generic clustering algorithm (e.g., K -means, with $K = N/10$) to initialize \mathbf{z} , and then initialize Φ accordingly. Then, in each iteration of MLP-B, we perform the following steps:

1. **Sample \mathbf{z} :** For each solution j , we remove it from its current cluster and sample its cluster assignment z_j from the posterior $p(z_j = k | \mathbf{z}_{-j}, \alpha, \mathbf{Y})$. Using Bayes rule, we have

$$p(z_j = k | \mathbf{z}_{-j}, \Phi, \alpha, \mathbf{Y}) = p(z_j = k | \mathbf{z}_{-j}, \phi_k, \alpha, \mathbf{y}_j) \propto p(z_j = k | \mathbf{z}_{-j}, \alpha) p(\mathbf{y}_j | z_j = k, \phi_k).$$

The prior probability $p(z_j = k | \mathbf{z}_{-j}, \alpha)$ is given by (4). For non-empty clusters, the observed data likelihood $p(\mathbf{y}_j | z_j = k, \phi_k)$ is given by (3). However, this does not apply to new clusters that are previously empty. For a new cluster, we

marginalize out ϕ_k , resulting in

$$\begin{aligned} p(\mathbf{y}_j | z_j = k, \beta) &= \int_{\phi_k} p(\mathbf{y}_j | z_j = k, \phi_k) p(\phi_k | \beta) \\ &= \int_{\phi_k} \text{Mult}(\mathbf{y}_j | z_j = k, \phi_k) \text{Dir}(\phi_k | \beta) \\ &= \frac{\Gamma(V\beta)}{\Gamma(\sum_{i=1}^V Y_{i,j} + V\beta)} \prod_{i=1}^V \frac{\Gamma(Y_{i,j} + \beta)}{\Gamma(\beta)}, \end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function.

If a cluster becomes empty after we remove a solution from its current cluster, then we remove it from our sampling process and erase its corresponding multinomial parameter vector ϕ_k . If a new cluster is sampled for z_j , then we sample its multinomial parameter vector ϕ_k immediately according to Step 2 below. Otherwise, we do not change ϕ_k until we have finished sampling \mathbf{z} for all solutions.

2. **Sample Φ :** For each cluster k , sample ϕ_k from its posterior $\text{Dir}(\phi_k | n_{1,k} + \beta, \dots, n_{V,k} + \beta)$, where $n_{i,k}$ is the number of times feature i occurs in the solutions that belong to cluster k .
3. **Sample α :** Sample α using the approach described in [41].
4. **Update β :** Update β using the fixed-point procedure described in [20].

The output of the Gibbs sampler is a series of samples that correspond to the approximate posterior distribution of the various parameters of interest. To make meaningful inference for these parameters (such as the posterior mean of a parameter), it is important to appropriately post-process these samples. For our estimate of the true number of clusters, \hat{K} , we simply take the mode of the posterior distribution on the number of clusters K . We use only iterations with $K = \hat{K}$ to estimate the posterior statistics [39].

In mixture models, the issue of “label-switching” can cause a model to be unidentifiable, because the cluster labels can be arbitrarily permuted without affecting the data likelihood. In order to overcome this issue, we use an approach reported in [39]. First, we compute the likelihood of the observed data in each iteration as $p(\mathbf{Y} | \Phi^\ell, \mathbf{z}^\ell)$, where Φ^ℓ and \mathbf{z}^ℓ represent the samples of these variables at the ℓ^{th} iteration. After the algorithm terminates, we search for the iteration ℓ_{\max} with the largest data likelihood and then permute the labels \mathbf{z}^ℓ in the other iterations to best match Φ^ℓ with $\Phi^{\ell_{\max}}$. We use $\hat{\Phi}$ (with columns $\hat{\phi}_k$) to denote the estimate of Φ , which is simply the posterior mean of Φ . Each solution j is assigned to the cluster indexed by the mode of the samples from the posterior of z_j , denoted by \hat{z}_j .

Auto-Grading via MLP-B

We now detail how to use MLP-B to automatically grade a large number N of learners’ solutions to a mathematical question, using a small number \hat{K} of instructor graded solutions. First, as in MLP-S, we select the set \mathcal{I}_B of “typical solutions” for the instructor to grade. We construct \mathcal{I}_B by selecting one

solution from each of the \hat{K} clusters that is most representative of the solutions in that cluster:

$$\mathcal{I}_B = \{\arg \max_j p(\mathbf{y}_j | \hat{\phi}_k), k = 1, 2, \dots, \hat{K}\}.$$

In words, for each cluster, we select the solution with the largest likelihood of being in that cluster.

The instructor grades the \hat{K} solutions in \mathcal{I}_B to form the set of instructor grades $\{g_k\}$ for $k \in \mathcal{I}_B$. Using these grades, we assign grades to the other solutions $j \notin \mathcal{I}_B$ according to

$$\hat{g}_j = \frac{\sum_{k=1}^{\hat{K}} p(\mathbf{y}_j | \hat{\phi}_k) g_k}{\sum_{k=1}^{\hat{K}} p(\mathbf{y}_j | \hat{\phi}_k)}. \quad (5)$$

That is, we grade each solution not in \mathcal{I}_B as the average of the instructor grades weighted by the likelihood that the solution belongs to cluster. We demonstrate the performance of auto-grading via MLP-B in the experimental results section below.

Feedback Generation via MLP-B

In addition to grading solutions, MLP-B can automatically provide useful feedback to learners on where they made errors in their solutions.

For a particular solution j denoted by its column feature value vector \mathbf{y}_j with V_j total expressions, let $\mathbf{y}_j^{(v)}$ denote the feature value vector that corresponds to the first v expressions of this solution, with $v = \{1, 2, \dots, V_j\}$. Under this notation, we evaluate the probability that the first v expressions of solution j belong to each of the \hat{K} clusters: $p(\mathbf{y}_j^{(v)} | \hat{\phi}_k), k = \{1, 2, \dots, \hat{K}\}$, for all v . Using these probabilities, we can also compute the expected credit of solution j after the first v expressions via

$$\hat{g}_j^{(v)} = \frac{\sum_{k=1}^{\hat{K}} p(\mathbf{y}_j^{(v)} | \hat{\phi}_k) g_k}{\sum_{k=1}^{\hat{K}} p(\mathbf{y}_j^{(v)} | \hat{\phi}_k)}, \quad (6)$$

where $\{g_k\}$ is the set of instructor grades as defined above.

Using these quantities, it is possible to identify that the learner has likely made an error at the v^{th} expression if it is most likely to belong to a cluster with credit g_k less than the full credit or, alternatively, if the expected credit $\hat{g}_j^{(v)}$ is less than the full credit.

The ability to automatically locate *where* an error has been made in a particular incorrect solution provides many benefits. For instance, MLP-B can inform instructors of the most common locations of learner errors to help guide their instruction. It can also enable an automated tutoring system to generate feedback to a learner as they make an error in the early steps of a solution, before it propagates to later steps. We demonstrate the efficacy of MLP-B to automatically locate learner errors using real-world educational data in the experiments section below.

EXPERIMENTS

In this section, we demonstrate how MLP-S and MLP-B can be used to accurately estimate the grades of roughly 100 open

Table 1: Datasets consisting of the solutions of 116 learners to 4 mathematical questions on algebra and signal processing. See the Appendix for the question statements.

	No. of solutions N	No. of features (unique expressions) V
Question 1	108	78
Question 2	113	53
Question 3	90	100
Question 4	110	45

response solutions to mathematical questions by only asking the course instructor to grade approximately 10 solutions. We also demonstrate how MLP-B can be used to automatically provide feedback to learners on the locations of errors in their solutions.

Auto-Grading via MLP-S and MLP-B

Datasets

Our dataset that consists of 116 learners solving 4 open response mathematical questions in an edX course. The set of questions includes 2 high-school level mathematical questions and 2 college-level signal processing questions (details about the questions can be found in Table 1, and the question statements are given in the Appendix). For each question, we pre-process the solutions to filter out the blank solutions and extract features. Using the features, we represent the solutions by the matrix \mathbf{Y} in (1). Every solution was graded by the course instructor with one of the scores in the set $\{0, 1, 2, 3\}$, with a full credit of 3.

Baseline: Random sub-sampling

We compare the auto-grading performance of MLP-S and MLP-B against a baseline method that does not group the solutions into clusters. In this method, we randomly sub-sample all solutions to form a small set of solutions for the instructor to grade. Then, each ungraded solution is simply assigned the grade of the solution in the set of instructor-graded solutions that is most similar to it as defined by \mathbf{S} in (2). Since this small set is picked randomly, we run the baseline method 10 times and report the best performance.³

Experimental setup

For each question, we apply four different methods for auto-grading:

- Random sub-sampling (RS) with the number of clusters $K \in \{5, 6, \dots, 40\}$.
- MLP-S with spectral clustering (SC) with $K \in \{5, 6, \dots, 40\}$.
- MLP-S with affinity propagation (AP) clustering. This algorithm does not require K as an input.
- MLP-B with hyperparameters set to the non-informative values $\alpha_\alpha = \alpha_\beta = 1$ and running the Gibbs sampling algorithm for 10,000 iterations with 2,000 burn-in iterations.

³Other baseline methods, such as the linear regression-based method used in the edX essay grading system [33], are not listed, because they did not perform as well as random sub-sampling in our experiments.

MLP-S with AP and MLP-B both automatically estimate the number of clusters K . Once the clusters are selected, we assign one solution from each cluster to be graded by the instructor using the methods described in earlier sections.

Performance metric

We use mean absolute error (MAE), which measures the “average absolute error per auto-graded solution”

$$\text{MAE} = \frac{\sum_{j=1}^{N-K} |\hat{g}_j - g_j|}{N - K},$$

as our performance metric. Here, $N - K$ equals the number of solutions that are auto-graded, and \hat{g}_j and g_j represent the estimated grade (for MLP-B, the estimated grades are rounded to integers) and the actual instructor grades for the auto-graded solutions, respectively.

Results and discussion

In Figure 5, we plot the MAE versus the number of clusters K for Questions 1–4. MLP-S with SC consistently outperforms the random sampling baseline algorithm for almost all values of K . This performance gain is likely due to the fact that the baseline method does not cluster the solutions and thus does not select a good subset of solutions for the instructor to grade. MLP-B is more accurate than MLP-S with both SC and AP and can automatically estimate the value of K , although at the price of significantly higher computational complexity (e.g., clustering and auto-grading one question takes 2 minutes for MLP-B compared to only 5 seconds for MLP-S with AP on a standard laptop computer with a 2.8GHz CPU and 8GB memory).

Both MLP-S and MLP-B grade the learners’ solutions accurately (e.g., an MAE of 0.04 out of the full grade 3 using only $K = 13$ instructor grades to auto-grade all $N = 113$ solutions to Question 2). Moreover, as we see in Figure 5, the MAE for MLP-S decreases as K increases, and eventually reaches 0 when K is large enough that only solutions that are exactly the same as each other belong to the same cluster. In practice, one can tune the value of K to achieve a balance between maximizing grading accuracy and minimizing human effort. Such a tuning process is not necessary for MLP-B, since it automatically estimates the value of K and achieves such a balance.

Feedback Generation via MLP-B

Experimental setup

Since Questions 3–4 require some familiarity with signal processing, we demonstrate the efficacy of MLP-B in providing feedback on mathematical solutions on Questions 1–2. Among the solutions to each question, there are a few types of common errors that more than one learner makes. We take one incorrect solution out of each type and run MLP-B on the other solutions to estimate the parameter $\hat{\phi}_k$ for each cluster. Using this information and the instructor grades $\{g_k\}$, after each expression v in a solution, we compute the probability that it belongs to a cluster $p(\mathbf{y}_j^{(v)} | \hat{\phi}_k)$ that does not have full credit ($g_k < 3$), together with the expected credit using (6). Once the expected grade is calculated to be less than full credit, we consider that an error has occurred.

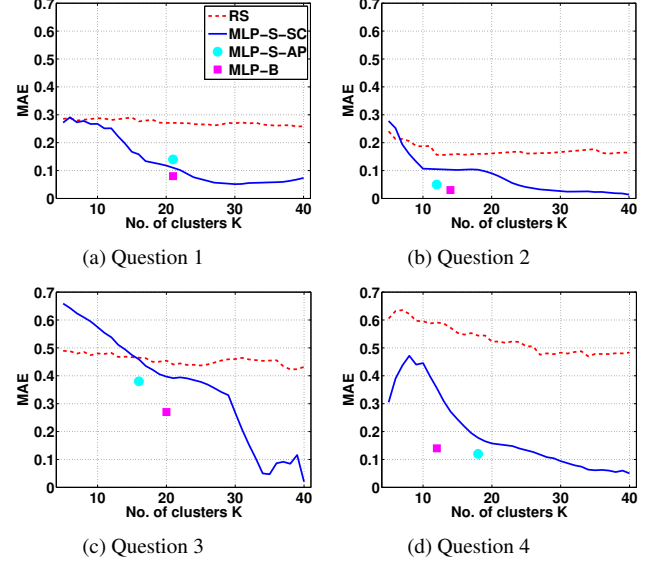


Figure 5: Mean absolute error (MAE) versus the number of instructor graded solutions (clusters) K , for Questions 1–4, respectively. For example, on Question 1, MLP-S and MLP-B estimate the true grade of each solution with an average error of around 0.1 out of a full credit of 3. “RS” represents the random sub-sampling baseline. Both MLP-S methods and MLP-B outperforms the baseline method.

Results and discussion

Two sample feedback generation process are shown in Figure 6. In Figure 6(a), we can provide feedback to the learner on their error as early as Line 2, before it carries over to later lines. Thus, MLP-B can potentially become a powerful tool to generate timely feedback to learners as they are solving mathematical questions, by analyzing the solutions it gathers from other learners.

CONCLUSIONS

We have developed a framework for mathematical language processing (MLP) that consists of three main steps: (i) converting each solution to an open response mathematical question into a series of numerical features; (ii) clustering the features from several solutions to uncover the structures of correct, partially correct, and incorrect solutions; and (iii) automatically grading the remaining (potentially large number of) solutions based on their assigned cluster and one instructor-provided grade per cluster. As our experiments have indicated, our framework can substantially reduce the human effort required for grading in large-scale courses. As a bonus, MLP-S enables instructors to visualize the clusters of solutions to help them identify common errors and thus groups of learners having the same misconceptions. As a further bonus, MLP-B can track the cluster assignment of each step of a multistep solution and determine when it departs from a cluster of correct solutions, which enables us to indicate the locations of errors to learners in real time. Improved learning outcomes should result from these innovations.

$$\begin{aligned}
& ((x^3 + \sin x)/e^x)' \\
&= (e^x(x^3 + \sin x)' - (x^3 + \sin x)(e^x)')/e^{2x} \\
\text{prob.incorrect} &= 0.11, \quad \text{exp.grade} = 3 \\
&= (e^x(2x^2 + \cos x) - (x^3 + \sin x)e^x)/e^{2x} \\
\text{prob.incorrect} &= 0.66, \quad \text{exp.grade} = 2 \\
&= (2x^2 + \cos x - x^3 - \sin x)/e^x \\
\text{prob.incorrect} &= 0.93, \quad \text{exp.grade} = 2 \\
&= (x^2(2 - x) + \cos x - \sin x)/e^x \\
\text{prob.incorrect} &= 0.99, \quad \text{exp.grade} = 2
\end{aligned}$$

(a) A sample feedback generation process where the learner makes an error in the expression in Line 2 while attempting to solve Question 1.

$$\begin{aligned}
& (x^2 + x + \sin^2 x + \cos^2 x)(2x - 3) \\
&= (x^2 + x + 1)(2x - 3) \\
\text{prob.incorrect} &= 0.09, \quad \text{exp.grade} = 3 \\
&= 4x^3 + 2x^2 + 2x - 3x^2 - 3x - 3 \\
\text{prob.incorrect} &= 0.82, \quad \text{exp.grade} = 2 \\
&= 4x^3 - x^2 - x - 3 \\
\text{prob.incorrect} &= 0.99, \quad \text{exp.grade} = 2
\end{aligned}$$

(b) A sample feedback generation process where the learner makes an error in the expression in Line 3 while attempting to solve Question 2.

Figure 6: Demonstration of real-time feedback generation by MLP-B while learners enter their solutions. After each expression, we compute both the probability that the learner's solution belongs to a cluster that does not have full credit and the learner's expected grade. An alert is generated when the expected credit is less than full credit.

There are several avenues for continued research. We are currently planning more extensive experiments on the edX platform involving tens of thousands of learners. We are also planning to extend the feature extraction step to take into account both the ordering of expressions and ancillary text in a solution. Clustering algorithms that allow a solution to belong to more than one cluster could make MLP more robust to outlier solutions and further reduce the number of solutions that the instructors need to grade. Finally, it would be interesting to explore how the features of solutions could be used to build predictive models, as in the Rasch model [27] or item response theory [18].

APPENDIX: QUESTION STATEMENTS

Question 1: Multiply

$$(x^2 + x + \sin^2 x + \cos^2 x)(2x - 3)$$

and simplify your answer as much as possible.

Question 2: Find the derivative of $\frac{x^3 + \sin(x)}{e^x}$ and simplify your answer as much as possible.

Question 3: A discrete-time linear time-invariant system has the impulse response shown in the figure (omitted). Calculate $H(e^{j\omega})$, the discrete-time Fourier transform of $h[n]$. Simplify your answer as much as possible until it has no summations.

Question 4: Evaluate the following summation

$$\sum_{k=-\infty}^{\infty} \delta[n-k] x[k-n].$$

Acknowledgments

Thanks to Heather Seeba for administering the data collection process and Christoph Studer for discussions and insights. Visit our website www.sparfa.com, where you can learn more about our project and purchase t-shirts and other merchandise.

REFERENCES

1. Attali, Y. Construct validity of e-rater in scoring TOEFL essays. Tech. rep., Educational Testing Service RR-07-21, May 2007.
2. Basu, S., Jacobs, C., and Vanderwende, L. Powergrading: A clustering approach to amplify human effort for short answer grading. *Trans. Association for Computational Linguistics* 1 (Oct. 2013), 391–402.
3. Blei, D., Griffiths, T., and Jordan, M. The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* 57, 2 (Jan. 2010), 7:1–7:30.
4. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *J. Machine Learning Research* 3 (Jan. 2003), 993–1022.
5. Brooks, M., Basu, S., Jacobs, C., and Vanderwende, L. Divide and correct: Using clusters to grade short answers at scale. In *Proc. 1st ACM Conf. on Learning at Scale* (Mar. 2014), 89–98.
6. Champaign, J., Colvin, K., Liu, A., Fredericks, C., Seaton, D., and Pritchard, D. Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In *Proc. 1st ACM Conf. on Learning at Scale* (Mar. 2014), 11–20.
7. Coursera. <https://www.coursera.org/>, 2014.
8. Cramer, M., Fisseni, B., Koepke, P., Kühlwein, D., Schröder, B., and Veldman, J. The Naproche project – Controlled natural language proof checking of mathematical texts, June 2010.
9. Dijkstra, J. A., and Khan, S. Khan Academy: The world's free virtual school. In *APS Meeting Abstracts* (Mar. 2011).
10. edX. <https://www.edx.org/>, 2014.
11. Frey, B. J., and Dueck, D. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.
12. Galenson, J., Reames, P., Bodik, R., Hartmann, B., and Sen, K. CodeHint: Dynamic and interactive synthesis of code snippets. In *Proc. 36th Intl. Conf. on Software Engineering* (June 2014), 653–663.
13. Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. *Bayesian Data Analysis*. CRC Press, 2013.

14. Griffiths, T., and Tenenbaum, J. Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems 16* (Dec. 2004), 17–24.
15. Gulwani, S., Radiček, I., and Zuleger, F. Feedback generation for performance problems in introductory programming assignments. In *Proc. 22nd ACM SIGSOFT Intl. Symposium on the Foundations of Software Engineering* (Nov. 2014, to appear).
16. Guo, P., and Reinecke, K. Demographic differences in how students navigate through MOOCs. In *Proc. 1st ACM Conf. on Learning at Scale* (Mar. 2014), 21–30.
17. Kang, S., McDermott, K., and Roediger III, H. Test format and corrective feedback modify the effect of testing on long-term retention. *European J. Cognitive Psychology 19*, 4–5 (July 2007), 528–558.
18. Lord, F. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates, 1980.
19. Megill, N. *Metamath: A computer language for pure mathematics*. CiteSeer, 1997.
20. Minka, T. Estimating a Dirichlet distribution. Tech. rep., MIT, Nov. 2000.
21. Naumowicz, A., and Kornilowicz, A. A brief overview of MIZAR. In *Theorem Proving in Higher Order Logics*, vol. 5674 of *Lecture Notes in Computer Science*. Aug. 2009, 67–72.
22. Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 2* (Dec. 2002), 849–856.
23. Nguyen, A., Piech, C., Huang, J., and Guibas, L. Codewebs: Scalable homework search for massive open online programming courses. In *Proc. 23rd Intl. World Wide Web Conference* (Seoul, Korea, Apr. 2014), 491–502.
24. OpenStaxTutor. <https://openstaxtutor.org/>, 2013.
25. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. Tuned models of peer assessment in MOOCs. In *Proc. 6th Intl. Conf. on Educational Data Mining* (July 2013), 153–160.
26. Raman, K., and Joachims, T. Methods for ordinal peer grading. In *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (Aug. 2014), 1037–1046.
27. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 1993.
28. Rivers, K., and Koedinger, K. A canonicalizing model for building programming tutors. In *Proc. 11th Intl. Conf. on Intelligent Tutoring Systems* (June 2012), 591–593.
29. Rivers, K., and Koedinger, K. Automating hint generation with solution space path construction. In *Proc. 12th Intl. Conf. on Intelligent Tutoring Systems* (June 2014), 329–339.
30. Sadler, P., and Good, E. The impact of self- and peer-grading on student learning. *Educational Assessment 11*, 1 (June 2006), 1–31.
31. Sapling Learning. <http://www.saplinglearning.com/>, 2014.
32. Singh, R., Gulwani, S., and Solar-Lezama, A. Automated feedback generation for introductory programming assignments. In *Proc. 34th ACM SIGPLAN Conf. on Programming Language Design and Implementation*, vol. 48 (June 2013), 15–26.
33. Southavilay, V., Yacef, K., Reimann, P., and Calvo, R. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proc. 3rd Intl. Conf. on Learning Analytics and Knowledge* (Apr. 2013), 38–47.
34. Srikant, S., and Aggarwal, V. A system to grade computer programming skills using machine learning. In *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (Aug. 2014), 1887–1896.
35. Steyvers, M., and Griffiths, T. Probabilistic topic models. *Handbook of Latent Semantic Analysis 427*, 7 (2007), 424–440.
36. SymPy Development Team. Sympy: Python library for symbolic mathematics, 2014. <http://www.sympy.org>.
37. Teh, Y. Dirichlet process. In *Encyclopedia of Machine Learning*. Springer, 2010, 280–287.
38. Vats, D., Studer, C., Lan, A. S., Carin, L., and Baraniuk, R. G. Test size reduction for concept estimation. In *Proc. 6th Intl. Conf. on Educational Data Mining* (July 2013), 292–295.
39. Waters, A., Fronczyk, K., Guindani, M., Baraniuk, R., and Vannucci, M. A Bayesian nonparametric approach for the analysis of multiple categorical item responses. *J. Statistical Planning and Inference* (2014, In press).
40. WebAssign. <https://webassign.com/>, 2014.
41. West, M. Hyperparameter estimation in Dirichlet process mixture models. Tech. rep., Duke University, 1992.
42. Wilkowski, J., Deutsch, A., and Russell, D. Student skill and goal achievement in the mapping with Google MOOC. In *Proc. 1st ACM Conf. on Learning at Scale* (Mar. 2014), 3–10.
43. Woolf, B. P. *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing E-learning*. Morgan Kaufman Publishers, 2008.
44. Yin, J., and Wang, J. A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (Aug. 2014), 233–242.