Effective Sampling for Large-Scale Automated Writing Evaluation Systems

Nicholas Dronen
University of Colorado & Pearson
Boulder, CO, USA 80301
dronen@colorado.edu

Peter W. Foltz
Pearson & University of Colorado
Boulder, CO, USA 80301
peter.foltz@pearson.com

Kyle Habermehl
Pearson
Boulder, CO, USA 80301
kyle.habermehl@pearson.com

ABSTRACT

Automated writing evaluation (AWE) has been shown to be an effective mechanism for quickly providing feedback to students. It has already seen wide adoption in enterprise-scale applications and is starting to be adopted in large-scale contexts. Training an AWE model has historically required a single batch of several hundred writing examples and human scores for each of them. This requirement limits large-scale adoption of AWE since human-scoring essays is costly. Here we evaluate algorithms for ensuring that AWE models are consistently trained using the most informative essays. Our results show how to minimize training set sizes while maximizing predictive performance, thereby reducing cost without unduly sacrificing accuracy. We conclude with a discussion of how to integrate this approach into large-scale AWE systems.

ACM Classification Keywords

I.2.7. Natural Language Processing: Text analysis; G.3. Probability and Statistics: Multivariate statistics

INTRODUCTION

Automated writing evalution (AWE) is the use of natural language processing and statistical modeling to score samples of writing. The first model for automatically scoring essays was created on punch cards in the 1960s [17]. At present there are several commercial automated essay scoring operations [19, 15, 1]. They are often used as the sole scorer in formative applications [25, 7] and increasingly in summative assessments (e.g. Pearson Test of English Academic). Even wider adoption of automated writing evaluation appears to be underway. Recently, for example, the Hewlettt Foundation sponsored open competitions for essay-length and short answer AWE modeling on Kaggle.com to help foster greater adoption of the technology [12, 13]. The edX platform for massive open online classes (MOOCs) contains a system for automated writing evaluation [6, 3, 5].

L@S 2015, March 14 - 18, 2015, Vancouver, BC

ACM 978-1-4503-3411-2/15/03$15.00
http://dx.doi.org/10.1145/2724660.2724661

Typically, an AWE model is trained using supervised learning and performs best when it has been trained for a specific prompt and scoring trait [9]. Training an AWE model starts with a batch of essays and one or more human scores for each essay. In the usual training procedure, the essays are preprocessed by a natural language processing pipeline, which results in a feature vector for each essay. The feature vectors and scores are then input to a learning algorithm that learns a model mapping feature vectors to scores. When the feature vector of a previously unseen essay is presented to the model, it can immediately provide a score.

Historically, the process of training an AWE model has required at least several hundred human-scored essays. The purpose of this requirement is twofold: to reduce sampling noise in the training set and to ensure that the model performs about as well as one trained with many more essays. In an online setting, however, a system for collecting essays and human scores can cause an AWE model to be trained whenever a new human score is entered into the system. For example, the edX AWE system allows models to be trained with as few as ten essays [4]. If the training set is sampled randomly, then training with so few essays will – with high probability – yield an AWE model that performs poorly.

The challenge of adopting AWE in large-scale contexts is that the best way to use the technology – with a customized scoring model for each prompt – is also quite expensive. In an enterprise environment, the human cost of scoring ranges from $3-6 USD per essay. Assuming 500 essays for a training set and 250 for test, the human cost per prompt is $2250-4500. In this paper we show how to overcome this key barrier to the widespread adoption of AWE in large-scale contexts. The method we use minimizes the number of essays that need to be scored and maximizes the information that each training example provides. This can also enable an online data collection system to intelligently choose which essays to be scored by humans. We conclude by discussing how to effectively integrate this method into AWE systems.

There are two bodies of literature about effective methods for choosing samples for training a supervised model. The older is *optimal experimental design* [8], a subfield of statistics that originated with the work of Kirstine Smith in the early 20th century [21]. The paradigmatic supervised learning algorithm for this literature is linear regression. In machine learning, *active learning* occupies much the same space as optimal de-

sign, although the literature tends to focus more on classification than regression [20].

The usual scenario for optimal experimental design is when samples are expensive to obtain, as with clinical trials or gathering samples during field studies in remote areas. An optimal design algorithm chooses samples from a pre-specified set of feature vectors in such a way as to minimize the variance in the supervised model. In active learning the assumption is usually that one begins with a set of *unlabeled* samples and wishes to obtain labels for them. For example, unlabeled samples can be HTML documents obtained by crawling the web and the desired labels can be the topics of the documents. Papers about regression in the active learning literature often cite the optimal design work of Valerii Fedorov (e.g. [16, 23, 2]).

When using an algorithm to choose samples prior to training a supervised learning algorithm, obtaining good results requires that the assumptions of the sampling algorithm match the assumptions of the learning algorithm. Here we use parametric linear regression and we employ an algorithm from the optimal experimental design literature [8] that is derived from the same foundations as linear regression.

Since we start with essays and ask which ones should be graded, ours is really the classic active learning scenario. We, however, stick mostly to the terminology from the optimal design literature throughout this paper.

DATA SETS

An AWE data set consists of some number of written responses and a set of scores for each response. In our experiments we use the industry track data from the Automated Student Assessment Prize (ASAP) Automated Essay Scoring (AES) competition sponsored by the Hewlett Foundation held in 2012 [12]. It contains training and test sets numbered 1 through 8. Key parts of the rubric of each set are summarized in Table 1.

In sets 1 and 2, a student is asked to write persuasively. Essay set 1 has the student take a position on the effects of computers on society and write a letter to the editor of a newspaper arguing for this position. Sets 3-6 are *response to source* prompts; they have the student read and analyze some source text and provide evidence from the text to support their analysis. The source of essay set 5 is a passage about the son of Cuban immigrants growing up in New Jersey. The student must describe the mood of the text. The narrative task tends to be more open ended. In essay set 8, the task is to write about an experience in which laughter was an important element.

The target variable of the data sets is the score or sum of scores that an essay was given by one or more human raters. The range of the target variable for each essay set is shown in Table 1. Essay set 2 has two target variables, which we identify as $2a$ and $2b$ depending on which target variable the model was trained to predict. For all essay sets the target variable includes all integers between the minimum and maximum scores. The target variable for essay set 7, for example, is the integers 2, 3, . . . , 30.

The size of the training and test sets are shown in Table 2. While relatively small, these test sets are nonetheless sufficient to ensure that the performance we report in the results section approaches the performance one might expect on a much larger sample. The same table shows the rounded average number of words in each essay set. A peculiarity of these sets is that in some cases the average length is quite low. This is true for all of the response-to-source sets, suggesting that students write less when performing a more focused task.

Before using an algorithm to select which essays raters should score, we compute the feature vector for each essay. In our experiments we use a subset of twenty-eight features – related to mechanics, grammar, lexical sophistication, and style – from the Intelligent Essay Assessor [9].

Table 1: ASAP AES rubric

Set	Task	Grade	Range
1	Persuasive	8	2-12
2a	Persuasive	10	1-6
2b	Persuasive	10	1-4
3	Source	10	0-3
4	Source	10	0-3
5	Source	8	0-4
6	Source	10	0-4
7	Narrative	7	2-30
8	Narrative	10	0-60

Table 2: ASAP AES data summary

Set	Train n	Test n	Avg. Words
1	1785	589	372
2a	1799	600	386
2b	1799	600	386
3	1699	568	110
4	1763	586	96
5	1805	601	124
6	1800	600	154
7	1469	495	174
8	917	304	636

REGRESSION MODELING

The target variable for automated essay scoring tasks is the student's score on an essay. It is typically an integer in a relatively narrow range determined by the scoring rubric. The ordering inherent in the target variable implies that essay scoring data are best modeled using a regression algorithm. Here we describe the approach we take to regression modeling in our experiments.

When the length of feature vectors in a design matrix ξ exceeds the number of feature vectors, the ordinary least squares solution is ill-posed. In this paper we allow m, the number of feature vectors to be as small as ten. Since p, the number of features, is twenty-eight, the ordinary least squares solution would be suboptimal for some of the models we train. Thus we use regularized regression instead of ordinary least squares. The regularized linear regression solution for β is the

solution for ordinary least squares augmented with a penalty function $J(\beta; \lambda)$,

$$\hat{\beta}(\lambda) = \arg\min_{\beta} [RSS(\beta) + J(\beta; \lambda)],$$

where λ is the regularization coefficient and is usually selected empirically (e.g. by cross validation).

We use ridge regression [10], which imposes a penalty on the L_2 norm of β, as in:

$$J(\beta; \lambda) = \lambda \sum_i \beta_i^2$$

This penalty constrains the coefficients such that irrelevant ones shrink towards zero without being eliminated from the model altogether.

Since the predictions of a linear model trained in such a way are real values, not integers, the predictions have to be mapped back onto the integer-valued score range. One method for determining the integer-valued score for a real-valued prediction $\hat{y}$ that lies between adjacent scores z_1 and z_2 is to apply thresholds as follows:

$$Score(\hat{y}, z_1, z_2) = \begin{cases} z_1 & \text{if } \hat{y} \leq \frac{z_1+z_2}{2}; \\ z_2 & \text{otherwise.} \end{cases} \quad (1)$$

Other reasonable approaches to the problem are to use ordinal logistic regression or to choose the thresholds based on performance on a validation set. Here we use Equation 1.

OPTIMAL DESIGN ALGORITHMS

In our experiments, we start with a set of n candidate essays. We wish to sample a subset of m of these essays that will give us a reasonably accurate predictive model after a human scores them.

More formally, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the n length-p feature vectors of the essays. Call this the *feature space*. The ith row of $\mathbf{X}$ is x_i and the jth column is x_j^T. Assume we lack the *target variable* $\mathbf{Y} \in \mathbb{R}^{n \times 1}$. Also let Ξ be the set of all subsets of rows of $\mathbf{X}$; this is the *design space*. A *design* ξ is an m-row member of Ξ. The goal of the algorithms described in this section is to find ξ such that, when it is agumented with $\mathbf{Y}_\xi$ – the $m \times 1$ matrix containing the target for the feature vectors in ξ – a supervised model that learns $P(\mathbf{Y}_\xi|\xi)$ has better predictive performance on unseen feature vectors than one would expect if ξ were chosen at random.

In our case the model $P(\mathbf{Y}_\xi|\xi)$ is linear regression. Consider a regression with feature x and target y. If we only allow an algorithm to choose two (in this case, unidimensional) feature vectors, which choices might result in a better regression model? The consequences of different choices are illustrated in Figure 1, where one can see that the greater the distance between the x values, the greater the probability that the model will perform almost as well as one trained with all of the data. This is the basic concept behind effective sampling of feature vectors for regression models. It motivates our selection of the algorithms we discuss in this section. The choices they make constitute a spectrum from choosing maximally distant feature vectors to choosing uniformly distant ones.

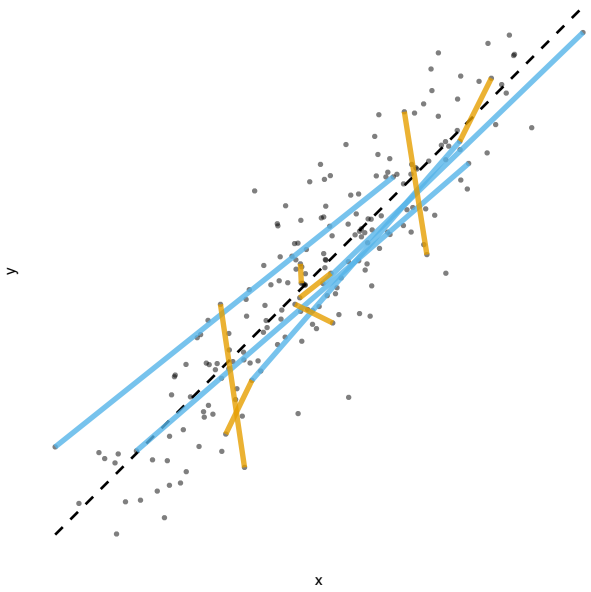


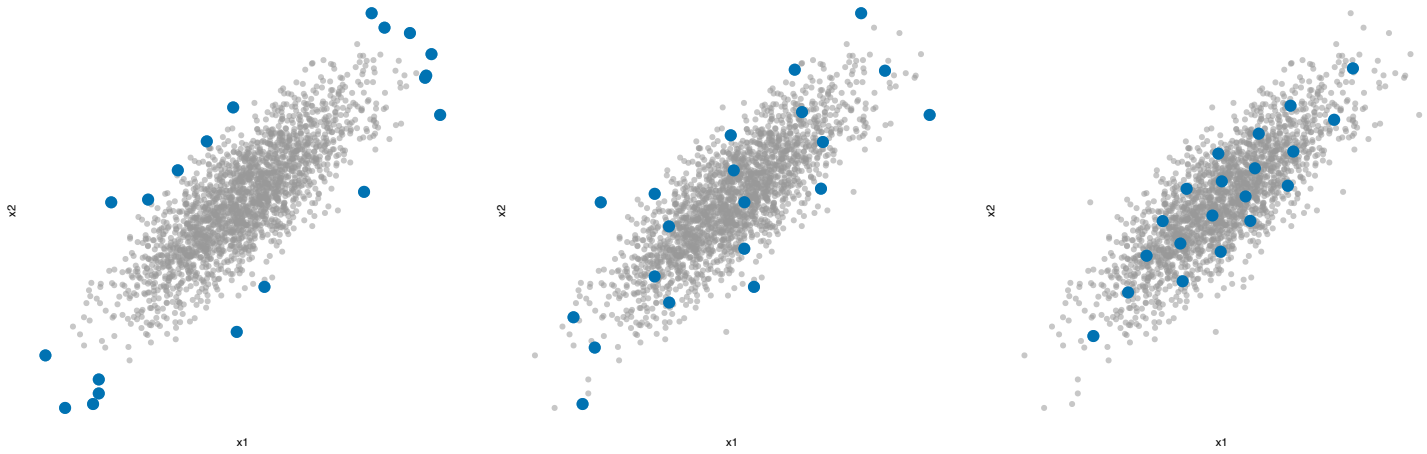
Figure 1: When fitting a regression line on a subset of the data, using distant points increases the probability that the fit will be close to the best fit. The regression line for all of the data is the dashed line. For contrast, lines connecting near and distant points are shown in different colors. The data are two hundred simulated points, with x drawn from a random normal distribution and $y = 1.5x + \epsilon$.

The R statistical language [18] has implementations of the algorithms we discuss in this section. The implementation of the Fedorov algorithm we used is in the AlgDesign package [24]. The prospectr package [22] contains implementations of Kennard-Stone and k-means sampling.

Fedorov exchange algorithm (with D-optimality)

The Fedorov exchange algorithm [8] is a greedy stepwise algorithm for finding ξ. Its purpose is to optimize various optimal design criteria – D-optimality, A-optimality, I-optimality, or others [11]. Most criteria are some function of the information matrix of ξ, defined as $M(\xi) = \frac{1}{m}\xi^T\xi$. D-optimality, for instance, searches for the ξ that maximizes the determinant $det\ M(\xi)$. In experiments with internal data sets, we didn't see major differences among D-optimality, A-optimality, or I-optimality, but the D-optimality criterion did tend to perform somewhat better, so for simplicity of exposition we limit ourselves to D-optimality in the experiments we conduct with the ASAP essay scoring data.

The first step of the Fedorov algorithm is to randomly initialize ξ with m rows from X. Subsequent steps replace a row in ξ with a row from X that is not already in ξ. The swapped rows are chosen greedily to optimize the design criterion. The algorithm can get stuck in a local optimum, depending on the initial choice of ξ. The remedy for this is to run the algorithm several times with different initial states and to choose the best one. The resulting design ξ consists of feature vectors



(a) Fedorov with D-optimality (b) Kennard-Stone (c) K-means

Figure 2: The points chosen by algorithms are distributed differently in the feature space. Here the larger points are twenty points selected by an algorithm. The data are simulated. In each figure, a point is one out of a thousand samples drawn at random from a two-dimensional multivariate normal distribution.

that are maximally distant from the centroid of the feature space, as can be seen in Figure 2a.

Kennard-Stone algorithm

The initialization step of the Kennard-Stone algorithm [14] is similar to D-optimality in that it chooses two feature vectors at the periphery of the feature space. Thereafter, however, it chooses feature vectors so as to distribute them uniformly. More precisely, let $d(u, v)$ be the distance between u and v. The first step of the Kennard-Stone algorithm is to choose indices i, j such that

$$\arg\max_{i,j \in 1 \ldots n} d(x_i, x_j),$$

then to initialize ξ with x_i, x_j and set $I_\xi = \{i, j\}$. Define

$$\Delta_{x_j}(\xi) = \min_{i \in I_\xi} d(x_j, \xi_i).$$

On each subsequent step the index j is added to I_ξ and x_j is added to ξ:

$$\arg\max_{j \notin I_\xi} \Delta_{x_j}(\xi).$$

After initialization, the point that is furthest from the point in the existing design are added at each step. The distances can be computed in any metric space. Here they are computed in Mahalanobis space. Figure 2b shows that a few of the feature vectors n the design are maximially distant from the centroid of the feature space and the rest are maximally distant from one another.

k-means sampling algorithm

This is simply the k-means clustering algorithm with an extra step at the end: after determining the final centroids, the design is determined by choosing the observation closest to each of the k centroids. Thus when searching for a design of size m, we set $k = m$.

Here the design is approximately uniformly distributed throughout the feature space in Figure 2c. This design does not include feature vectors at the periphery of the feature space, because the centroid of a cluster to which a peripheral point belongs will almost always be closer to some other point that is not peripheral. The uniformity on display here is in part an artifact of the simulated data. In real data the feature vectors would be distributed less than smoothly, so the centroids of the final clusters would be distributed less than uniformly.

Persistence of selections

An important question is the degree to which the feature vectors selected by these algorithms persist across sequential values of m. If the design ξ at $m - 1$ is always included in ξ at m, its selections persist perfectly. The degree of persistence can be measured as

$$Persistence = \frac{|\xi_m \cap \xi_{m+1}|}{|m|}. \quad (2)$$

The persistence of selections for increasing values of m is shown in Figure 3. Kennard-Stone starts with the same two maximally distant feature vectors and adds further ones sequentially according to a deterministic procedure, so its selections are perfectly persistent. The Fedorov algorithm runs several times with different starting points in order to avoid getting stuck in local optima. This results in some loss of persistence, which appears to improve as m increases. In the range $150 < m < 175$, however, the Fedorov algorithm's persistence is erratic, suggesting the existence of multiple local optima that are nearly indistinguishable according to the optimality criterion. The k-means algorithm exhibits the lowest persistence of the three algorithms. This is expected, as the final centroids will change as k, the number of clusters, increases.

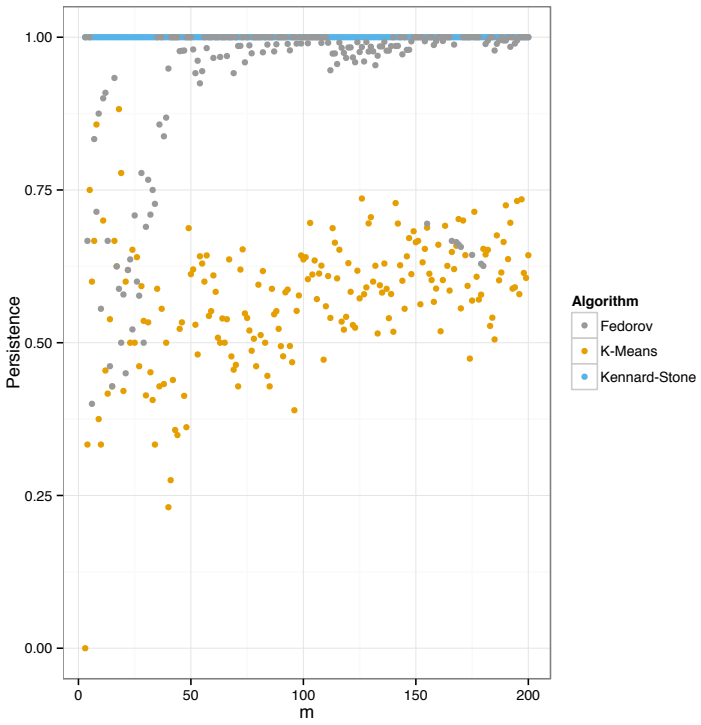


Figure 3: The persistence of selections from ASAP AES essay set 1. The Kennard-Stone algorithm's persistence is 1 everywhere.

EVALUATION

To evaluate the performance of the algorithms, the ideal would be to have access both to an effectively infinite number of unscored essays and to an oracle that could provide scores upon request. Since we have finite data, we simulate a much larger amount of data by repeatedly sampling half of the entire training set (without replacement) before choosing feature vectors to be scored by humans. Of course, since we already have the scores, the human-scoring step of our evaluation doesn't actually occur. The algorithms don't look at the scores, however, so no information leaks unfairly into the process.

We let the desired training set size m range from 10 to 100 in increments of 10. This allows us to simulate an operational environment where new essays arrive over time. After fixing m and sampling half of the entire training set, we use each algorithm to choose a subset of m feature vectors, pretend to have the corresponding essays scored by humans, and train one regression model with each algorithm's chosen subset. A model's performance on the test set is the evaluation criterion – specifically, the Pearson correlation coefficient of its rounded predictions with the scores of the human raters. Since test set essays are the same for all models, the performance of each model is comparable regardless of the size of the training set. To provide smoothed estimates, we averaged the correlations for each of the 300 models trained using a particular combination of training set, sampling algorithm, and size m.

The baseline design algorithm is random sampling without replacement. For a given design size m, the baseline algorithm selects a design ξ_m of unique feature vectors from X at random. Models trained using a design chosen by an optimal design algorithm are expected to perform better than ones trained using this baseline. Assuming an algorithm performs better than the baseline, the key question is how its performance over the baseline changes as m increases. The hallmark of an effective optimal design algorithm is that it prefers to choose the essays that are most informative to the model; this should manifest itself as a *larger* margin above the baseline when m is small. As $m \to n$ (where n is the number of feature vectors from which to choose), the margin should go to 0.

RESULTS

Table 3 shows the percent change of the average of each algorithm's models over the baseline. To test for significance, we transform the correlations using the Fisher Z-transform before applying the two-sample t-test at the 0.001 level. The test is an approximation, as the normality assumption of the t-test does not hold, but the correlation coefficients are unimodal and the sample sizes are relatively large, so we assume that the approximation is quite good. For interpretability, the percent change is calculated using the Pearson correlation. The means of Fedorov and Kennard-Stone results are significantly different from the baseline for most ASAP AES essay sets and training set sizes Fedorov and Kennard-Stone, with the exception of the anomalous essay set 4. The k-means results are more often indistinguishable from the baseline.

To achieve parity with the baseline at $m = 100$, the Fedorov algorithm typically needs to select only about 50 essays. With essay set 1, for instance, it only needs to select 10 essays, as can be seen in Figure 4.

With respect to average performance, the Kennard-Stone algorithm tends to perform well on the same sets as the Fedorov algorithm. The confidence intervals in Figure 4 aren't prominent enough to illustrate a notable difference. The standard deviations of the Kennard-Stone models tend to be greater, and for some test sets are even worse than the baseline, as can be seen in Figure 5. This suggests that the Fedorov algorithm is the most robust: models trained with the essays it selects have the best average performance and the least variance. It also supports the notion – illustrated in for 1-d data in Figure 1 – that the periphery of the feature space is the most effective region from which to choose feature vectors.

We should expect the Fedorov algorithm to perform best when the underlying process that generates the data is best fit using a linear model. Overall, the ranking of optimal design algorithms – from best to worst – conforms to this expectation: Fedorov, Kennard-Stone, and k-means. The only anomaly is essay set 4, where the k-means models perform best for $10 < m < 60$, with respect both to the mean of the Pearson correlation coefficient and its standard deviation. Two possible explanations of this are noise introduced by

Set	m									
	10	20	30	40	50	60	70	80	90	100
1	42	18	10	7	5	4	3	2	2	2
2a	86	69	47	37	26	26	22	18	18	15
2b	94	68	48	39	32	27	26	23	18	17
3	26	18	16	10	8	8	5	5	4	4
4	25	7	**4**	**3**	5	4	**3**	**3**	**4**	**2**
5	25	12	7	4	3	2	1	1	1	1
6	59	30	17	12	8	7	5	4	3	3
7	54	29	16	10	6	5	4	2	2	2
8	106	84	62	93	80	65	74	84	50	63

(a) Fedorov

Set	m									
	10	20	30	40	50	60	70	80	90	100
1	36	15	9	6	5	4	3	2	2	2
2a	62	70	48	36	26	26	21	18	18	15
2b	70	67	45	36	30	25	25	22	17	17
3	25	13	13	9	7	9	6	5	5	4
4	-33	**-2**	**-1**	**0**	**1**	**2**	**1**	**2**	**2**	**1**
5	**-5**	11	7	5	4	3	2	1	1	1
6	52	28	17	11	8	6	4	4	2	3
7	25	22	11	8	5	4	3	2	3	2
8	52	66	45	54	73	53	69	85	53	62

(b) Kennard-Stone

Set	m									
	10	20	30	40	50	60	70	80	90	100
1	-8	**-1**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
2a	30	42	32	23	17	19	16	13	14	12
2b	53	40	32	28	23	19	20	18	14	14
3	**-4**	**0**	5	**3**	**1**	3	1	1	1	1
4	13	14	13	7	8	6	4	4	**3**	**0**
5	**4**	9	5	4	3	2	1	1	1	1
6	15	8	10	9	7	6	5	4	3	3
7	**0**	7	8	6	4	4	3	2	2	2
8	46	52	45	41	40	33	40	43	**22**	25

(c) K-means

Table 3: Percent change over baseline of the mean performance of models trained with essays selected by optimal design algorithms. Statistically insignificant differences are shown in bold. Significance is determined by a two-sample t-test of means of z-transformed correlations with a 0.001 significance level.

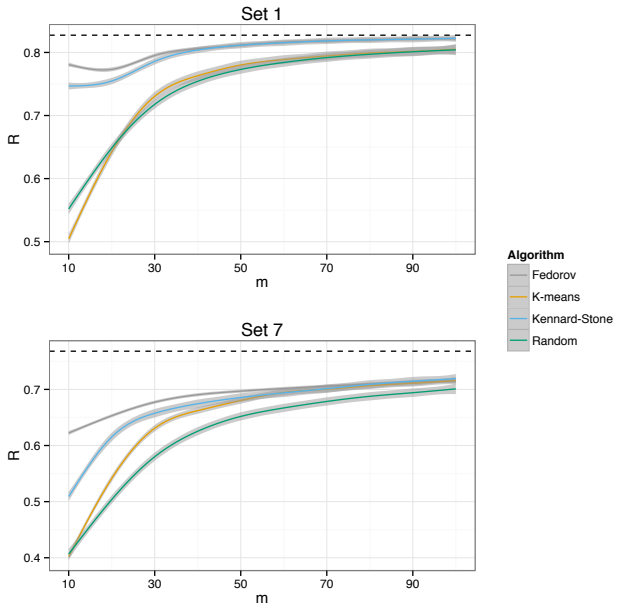


Figure 4: The performance of each algorithm on two test sets. Lines are the smoothed average of 300 models with 95% confidence intervals of the mean. The X axis, m, is the number of essays selected. The Y axis, R, is the Pearson correlation between the model and human raters on the test set. The dashed horizontal line is the mean test set performance when training with all of the data available *in an iteration*.

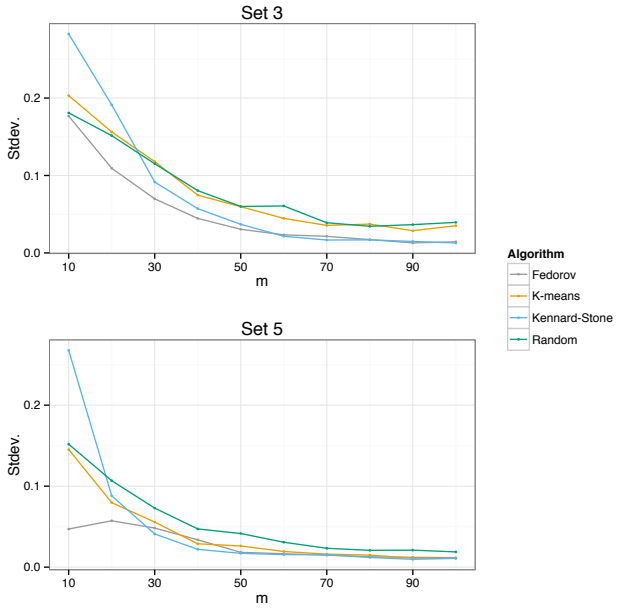


Figure 5: Standard deviations of each algorithm's performance on two test sets. The Y axis is the standard deviation of the Pearson correlation coefficients of 300 models trained with essays selected by an algorithm. The X axis, m, is the number of essays selected.

human raters or an insufficient set of features in the feature space.

Based on the relative performances of the three optimal design algorithms, we strongly favor the use of the Fedorov algorithm with D optimality. The Kennard-Stone algorithm's performance is a close second, but the models trained with the essays it selects tend to have greater variance and when m is small it doesn't approach the impressive performance of the Fedorov algorithm.

CONCLUSION

Effective sampling disciplines the unruly process of obtaining training sets for AWE models. Our simulations show that models trained using a relatively small number of intelligently-chosen essays tend to perform well on out-of-sample essays, at times almost as well as a model trained on an entire training set. In several of the data sets, training with 30-50 essays yields approximately the same performance as a model trained with hundreds of essays. The key implication is that it is possible to minimize the number of human scores necessary for training an AWE model without unduly sacrificing accuracy.

These results also have implications for designing systems that allow rapid integration of new AWE scoring prompts into large-scale systems. One such approach is to train models incrementally. For example, in a large MOOC, a professor may post a new writing prompt for the students. After the first few hundred students submit their essay, the optimal design algorithm selects a subset for the professor or teaching assistant to score. Using this initial training set, a scoring model can quickly be built that will start automatically scoring student essays, so that students can receive feedback on their writing. As more essays are received from students, the optimal design algorithm can select additional essays for human scoring that will be most informative to help refine the accuracy of the scoring model. The model can then be automatically retrained with even greater accuracy of scoring.

When a model is retrained as scores arrive into a data collection system, using optimal design to select the essays from a large pool will ensure that the initial training set consists only of essays that are likely to result in a reasonable model. It will also both ensure that subsequently-added essays are maximally informative and minimize the duplication of the effort of human raters.

Optimal design is necessary but not sufficient for making incremental training of AWE models effective and accurate. While it minimizes the sampling noise in a training set, it doesn't guarantee a minimum level of performance. To ensure that models are not deployed prematurely, a validation set of essays should first be selected from the pool and scored by humans. Measuring model performance with a validation set will ensure that the model performs well enough before it is deployed. Each time a new human score enters the data collection system, the model can be retrained, and its performance on the validation set can be computed. If the model doesn't perform well enough to be deployed to a production environment, additional essays can be selected to be scored. This process should repeat until the model reaches an acceptable level of performance.

In experiments with internal data sets, we noticed that when a model trained on the entire training set performed to our satisfaction, optimal design performed well too. The conditions that caused a model trained on the entire training set to perform poorly also caused the performance of optimal design to degrade. The most common cause of not meeting the conditions for good model performance is noise due either to poor agreement among human raters or model misspecification. An example of model misspecification is when the feature set does not capture textual semantics and the rubric instructs the human raters to give a higher score to an essay if it covers some set of topics[1]. Our recommended solution to avoiding model misspecification, which we don't evaluate here, is to be generous with one's feature set and to let a regularizer shrink uninformative features. We expect the results we report here to generalize to other tasks and populations as long as the conditions are right for a model trained on the entire training set to perform satisfactorily.

While these results are promising for the scoring of essays, we expect that it will be less effective with short answer data sets. Short answer data sets are often modeled using a learning algorithm that employs a set of decision trees, such as random forest. The feature vectors are often different, too, because such models perform well when they know exactly which words a student used. Consequently, it is common for the feature vectors to be rows of a *document-term matrix* – an occurrence matrix with terms as columns and documents as rows – constructed from the training set responses. In that representation, the feature vectors are sparse and relatively long. Optimal design and active learning work best when the assumptions of the sampling algorithm agree with the assumptions of the subsequent supervised learning algorithm. Thus, using the Fedorov algorithm to select a subset of short responses to be scored by humans when the supervised model is random forest trained with a document-term matrix is not guaranteed to yield the best results.

The use of optimal design for AWE training sets can be seen as a form of manufacturing process control in which the amount of information provided by a human score is maximized and its variance minimized. From this perspective, the goal of our use of optimal design is primarily to make the output of the human scoring process – that is, the quantity of information in a single human score – more predictable. The effect of this is to enable safe and reliable incremental training of AWE models. Overall, the approach solves a barrier to the adoption of AWE into large-scale formative and summative systems, allowing the computer to minimize the amount of human effort needed to collect and score essays by choosing effectively which essays humans need to score.

[1]There has been some recent work on active learning for misspecified models [23, 2], but their thrust has been primarily theoretical and do not appear to be immediately helpful for this problem.

REFERENCES

1. Attali, Y., and Burstein, J. Automated essay scoring with e-rater v. 2. *The Journal of Technology, Learning and Assessment 4*, 3 (2006).
2. Bach, F. R. Active learning for misspecied generalized linear models. *Adv. Neural Inf. Process. Syst.* (2007).
3. `https://github.com/edx/discern`. Accessed: 2013-10-7.
4. `https://github.com/edx/discern/blob/master/ml_grading/ml_model_creation.py`. Accessed: 2013-10-7.
5. `https://github.com/edx/ease`. Accessed: 2013-10-7.
6. `http://edx.org`. Accessed: 2013-10-7.
7. `https://criterion.ets.org`. Accessed: 2013-10-13.
8. Fedorov, V. V. *Theory Of Optimal Experiments*. Probability and mathematical statistics. Elsevier Science, 1972.
9. Foltz, P. W., Streeter, L., Lochbaum, K. E., and Landauer, T. K. Implementation and applications of the intelligent essay assessor. In *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. Taylor & Francis, 2013, 68–88.
10. Hoerl, A. E., and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1 Feb. 1970), 55–67.
11. John, R. C. S., and Draper, N. R. D-Optimality for regression designs: A review. *Technometrics 17*, 1 (1975), 15–23.
12. The hewlett foundation: Automated essay scoring — kaggle. `http://www.kaggle.com/c/ASAP-AES`. Accessed: 2013-10-7.
13. The hewlett foundation: Short answer scoring — kaggle. `http://www.kaggle.com/c/ASAP-SAS`. Accessed: 2013-10-7.
14. Kennard, R. W., and Stone, L. A. Computer aided design of experiments. *Technometrics 11*, 1 (1 Feb. 1969), 137–148.
15. Landauer, T. K., Laham, D., and Foltz, P. W. Automatic essay assessment. *Assessment in Education: Principles, Policy and Practice 10*, 3 (2003), 295–308.
16. MacKay, D. Information-Based objective functions for active data selection. *Neural Computation 4*, 4 (July 1992), 590–604.
17. Page, E., and Paulus, D. The analysis of essays by computer. final report.
18. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.
19. Rudner, L. M., Garcia, V., and Welch, C. An evaluation of IntelliMetric essay scoring system. *The Journal of Technology, Learning and Assessment 4*, 4 (2006).
20. Settles, B. Active learning literature survey. *University of Wisconsin, Madison* (2010).
21. Smith, K. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika 12*, 1/2 (1 Nov. 1918), 1–85.
22. Stevens, A., and Ramirez-Lopez, L. *An introduction to the prospectr package*, 2013. R package version 0.1.3.
23. Sugiyama, M. Active learning for misspecified models. *Adv. Neural Inf. Process. Syst.* (2005).
24. Wheeler, B. *AlgDesign: Algorithmic Experimental Design*, 2014. R package version 1.1-7.2.
25. `http://writetolearn.net`. Accessed: 2013-10-13.