Improving performance of automated scoring through detection of outliers and understanding model instabilities

Research Report

Peter W. Foltz

Mark Rosenstein

Karen Lochbaum

Foltz, P. W., Rosenstein, M., & Lochbaum, K. E. (2013). Improving performance of automated scoring through detection of outliers and understanding model instabilities. Paper presented at *the National Council on Measurement in Education Conference (NCME*), San Francisco, CA, April.

April 2013

Abstract

Automated scoring of writing operates under the assumption that the quality of a student essay can be characterized by a model of weighted features that are extracted from the essay. Because the automated scoring model is trained predicated on a representative sample of essays for a prompt, associated with each feature is an expected range of acceptable values based on the distribution in the training set. However, if the value for a particular feature or combined values for a group of features extend beyond the training range, the assumptions of the scoring model may be violated and may cause instabilities in the model. This issue is often compounded by utilizing features that are not distributed normally or that exhibit a nonlinear relationship to scores. This talk will describe a series of analyses which use methods that examine the impact on scoring of essays at or beyond the boundaries exhibited in the training set. The talk will further illustrate how these approaches contribute to developing scoring confidence measures and the detection of outlier essays that could be due to construct-irrelevant responses. Such scoring confidence measures can be integrated into a combined human/computer-based scoring scheme in which essays with low computer-scoring confidence can be passed on for human judgment.

Improving performance of automated scoring through detection of outliers and understanding model instabilities

Automated scoring of essays offers a means to judge the quality of student essays with rapid feedback to teachers, students and decision-makers. While automation allows nearly immediate and often cost-effective scoring, it is critical to ensure that that the techniques used also produce accurate, reliable, and valid scoring for the set of essays that will be input to the system (e.g., Williamson et al., 2010). However, designers of automated techniques cannot always anticipate the full diversity of inputs they will receive. The models may be asked to score essays that are highly unusual or not representative of the expected input. Thus, it becomes incumbent on developers of scoring systems to ensure that not only are these systems able to accurately score a wide range of input essays, but they are also able to judge the potential of any incoming essay to not be accurately scored.

Techniques for automated scoring of writing typically measure the quality of an essay by extracting a set of features from the essay and then combining the features using a statistical model that is based on a training set of essays. In developing a scoring system, certain assumptions are made about the distribution of the essays used for training the system as well as the statistical characteristics of the features and the modeling techniques used. With careful consideration of the assumptions being made, it is possible to analyze information from essay features and modeling techniques to develop methods that provide levels of confidence that an essay can be scored accurately. In addition, these approaches may permit a better understanding of how different essay features contribute to an essay's score and how the features may operate under different modeling assumptions that can affect scoring. In this paper we provide a series of analyses which examine the impact of essay features and model assumptions on scoring and describe how these can be used to improve the detection of essays that can not be scored accurately.

Outlier detection considered within the automated scoring process

Throughout the development of automated scoring systems and any item specific automated scoring model, there are a number of stages that influence both the reliability of the scoring as well as help determine the suitability of a given essay for automated scoring. These stages include:

1. The collection of essays used to train the scoring system.
2. The collection of scores from human raters for the training essays.
3. The creation and testing of features and algorithms that most reliably detect components of student writing quality and knowledge.
4. The use of methods that detect essays that may be scored less reliably by the automated scoring methods in the implemented system.

While we have previously focused on how these stages impact reliability (see Foltz et al., 2012), here we consider how aspects of the scoring process impact detection of outlier essays that may be scored less accurately.

The collection of essays used for the training set provide some boundaries for determining what will be considered appropriate input for scoring. Essays used in the training set should reflect distributional properties of the expected essays, and generally, the scores for the training essays should represent a normal distribution, while ensuring that there are sufficient (e.g., at least a minimum of 10-20) examples at each score point. (See Foltz et al., 2012). In addition, the training essays should be representative of the expected set of essays that will be scored including considerations of the topics covered, length, type of language used in the essays as well as sampled appropriately from the expected student population for gender, ethnicity, and skill levels. (e.g., Williamson et al., 2010). This training set, therefore provides a characterization of the expected range of essays that will be received for scoring.

The features used to analyze the essays provide a means to ascertain how well any essay falls within the distributional confines of the training set. These features can include measuring aspects of the writing such as the caliber of the student's expression and organization of words and sentences, the student's knowledge of the domain content, the quality of the student's reasoning, and the student's skills in language use, grammar and mechanics of writing. Generally, these features need to be construct relevant, provide effective measures of the target skill, and are well represented in the expected population of essays to be scored. Building a scoring model based on these features makes implicit assumptions that the features will be represented across the spectrum of essays and that they will behave in a sufficiently regular manner to allow successful modeling of their behavior.

The algorithms, or modeling techniques, similarly make assumptions about features that will be used. For example, in linear regression-based approaches, the models posit that features will behave in a linear fashion within the range of expected essays, even when they are considered within a multivariate context. Such approaches also tend to assume that features will have normal distributions however, many language features have non-normal distributions, such as the Zipf distribution for word frequency. Non-linear and Bayesian-based approaches make analogous types of assumptions within the constraints of the range of their training sets. In each of these cases, it is assumed that variables falling within the expected range will provide stable estimates. However, if variables fall outside of the range, it is not always clear whether the algorithms will provide stable estimates and may degrade scoring accuracy.

By analyzing the features of the essays from the training set, we can therefore derive an expected range of essay features. The system can then determine if the value for a particular feature or the combined values for a group of features are beyond the training range. Those that fall outside the confidence interval of this range may indicate potential violations of the assumptions of the scoring model and may cause instabilities in the model. Below we examine approaches to utilizing these assumptions to detect outlier essays and essays that fall in a part of the feature space where the model may be less stable.

We illustrate three approaches to detection of outliers and the effects of model assumptions across the values of different scoring features. All the presented analyses were performed using data from a single, advanced high-school/entry level college prompt with 559 essays. The median essay length for the prompt was 389 words with an interquartile range of 173 and the median sentence count was 19 with interquartile range of 9. The approach generalizes to other prompts and essay sets, but is illustrated through analysis of features of a single prompt for concreteness and simplicity. Not all of the phenomena discussed are evidenced in every prompt and it is important to note that while this paper provides examples of a variety of features used to measure performance in essays, not all features described in this paper are used operationally in the Intelligent Essay Assessor for scoring. Thus, the focus of the paper is on illustrating how such generalized approaches can be used across a range of different types of features for detecting essays that deviate from the norm rather than any specific instantiation used in the IEA.

Multivariate Normality as a test for outliers

An essay that may be an outlier can be classified by it values on a single feature or it could be the conjoint values across multiple features. Indeed, for the case of multiple features, the value of each individual feature may appear within an acceptable range, but from a multivariate perspective, the combination of features may fall in a part of the feature space were there is little or no data from the training set. In such areas of sparsity, it is less certain that that the values of the features will be combine in a way that will be representative of the best estimate of the scores for the essay

It is often convenient to group the features describing responses into functional sets, which allows harnessing the covariance structure of each set to guide an evaluation of how far a given response is from responses in the training set. Comparability of that functional set to those from the training set provides a measure of confidence that the response is sufficiently represented by the responses within the training set to be accurately scored and conversely to signal responses inappropriate for a given scoring model. A distance measure within feature space, such as a generalized distance, is required to allow these types of comparisons. While generalized distance measures, such as the Mahalanobis Distance (1936), do not require distributional assumptions, the interpretation of the resulting distances (i.e. how large a separation should raise concern) is more straightforward if the underlying distribution is multivariate normal. This approach allows detecting the degree to which feature sets are multivariate normal. For many of our feature sets, this assumption is violated, where we see long tails as the most common deviation from normality.

Many statistics are available to help judge multivariate normality, such as the Cox-Small test (1978), though for a better understanding of the underlying phenomena, we tend to favor visualization. The asymptotic Chi-squared distribution of the Mahalanobis distances of data generated under multivariate normality, allows us to follow Healy (1968) in producing and examining the multivariate normal plot. This type of plot may feel familiar since it is quite similar to the quantile-quantile plots that are commonly used in comparing distributions of univariate data.

The following plots show the generalized distances from the same set of training responses, while varying the feature sets. The distances are plotted against the quantiles of the Chi-squared distribution. Deviations from the 45-degree line indicate departure from multivariate normality. The plots also include a bootstrap 95% confidence interval, indicated in red, to help guide interpretation.

The two plots in Figure 1 provide a baseline example of the behavior of generalized distance plots with the left plot indicating what a multivariate normal plot based on drawing from a 20 dimensional multivariate normal distribution of variables looks like. This pattern is then contrasted in the right plot by drawing from a multivariate t distribution with 10 degrees of freedom, which is slightly overdispersed from the normal. Clearly, the 95% confidence interval for this plot excludes the diagonal for much of the range in the right plot. This illustrates the basis to detect deviations from multivariate distributions.

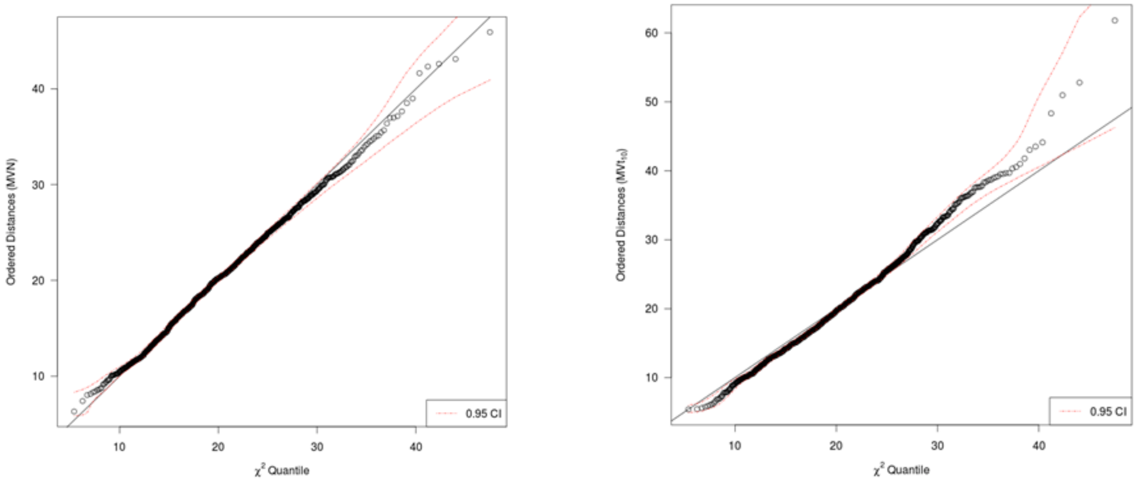

Figure 1a and b: Multivariate Normal plot. 1a uses random samples from a multivariate normal distribution. 1b uses samples drawn from a multivariate t distribution with 10 degrees of freedom. Note that the samples in 1b cause deviations of confidence interval from the 45 degree line.

The next three plots show examples taken from feature sets that are relevant for scoring: 1) coherence features, which measure aspects of the flow across sentences as well as how well each sentence contributes to the coherence of the overall essay (see Foltz, Kintsch & Landauer, 1998), 2) features derived from statistical language models (based on n-gram features (e.g., Jurafsky & Martin, 2009), and finally 3) a set of readability features. In all cases, except for the readability features multivariate normality is violated.

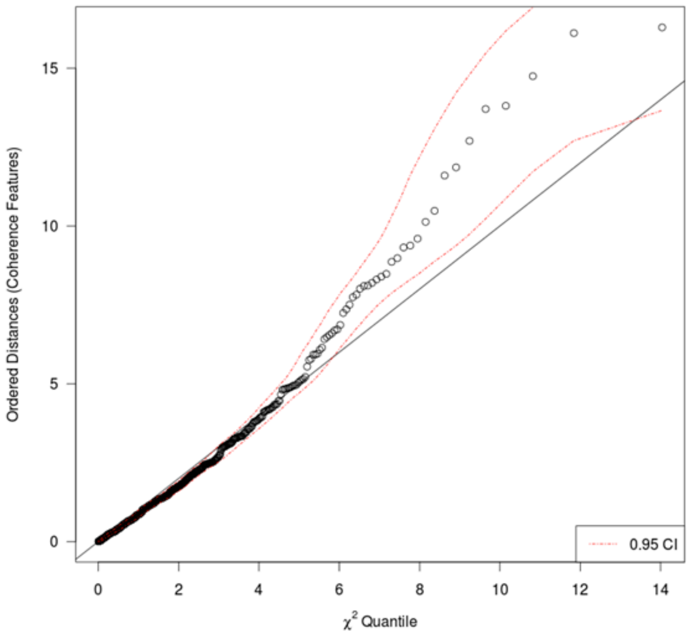

Figure 2: Multivariate normal plot of coherence features

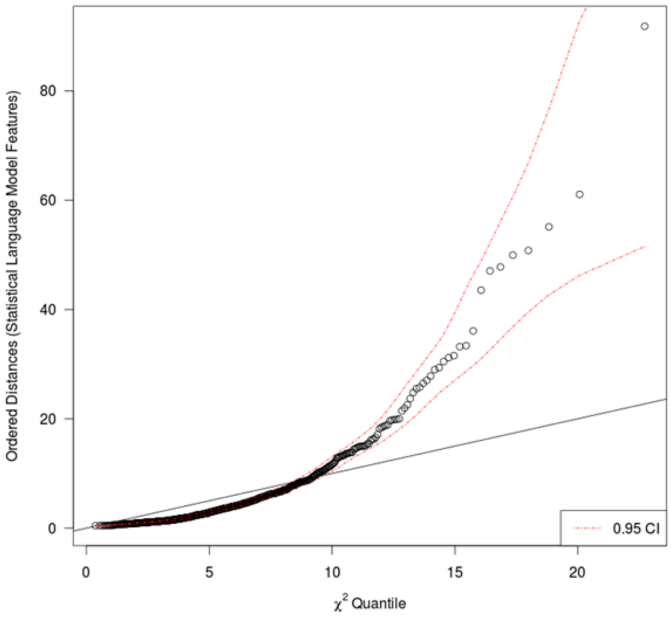

Figure 3: Multivariate normal plot of statistical language features

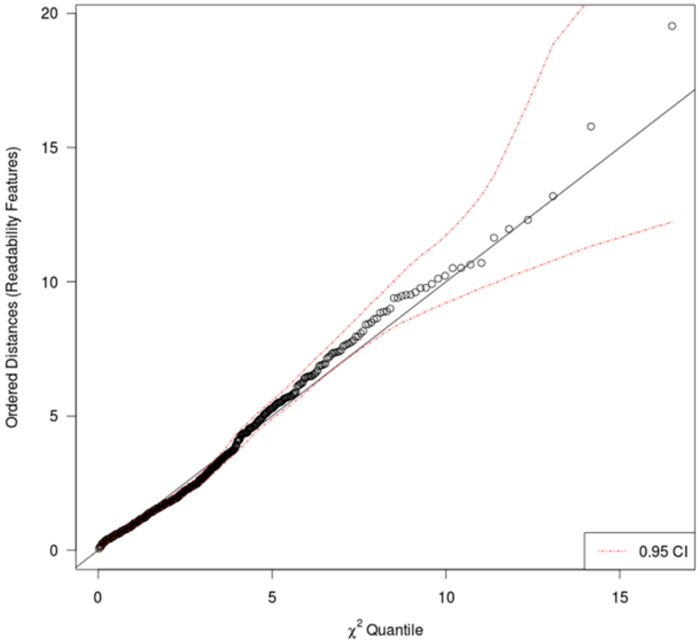

Figure 4: Multivariate normal plot of readability features

At least two lessons can be drawn from these examples, with the first being that confidence intervals for responses for these feature sets based on a multivariate normality assumption are going to be conservative. A significant number of responses that almost certainly could have been accurately scored will be falsely identified as outliers, which has a cost impact on automated scoring if the identified responses are then sent to human scorers when in fact automated scoring would have been sufficient. The second lesson is that to move away from conservative criteria requires a larger set of responses to allow more accurate estimates for the tails of these distributions.

Nonlinearities in features

From a theoretical perspective, it is not unexpected that empirically the relationship between human score and many of the features characterizing the responses are often not linear. Given the fortunate bounty of linear approximations over much of the range of these features, a linear approximation will often fit quite well. It is generally only at the extremes that we see notable deviation from linearity, though these cases commonly cause the most consternation. Beyond deviations at the extremes, the next most frequent nonlinearity arises in features that have a delimited range of optimal values, with human scores decreasing on both sides away from the optimal values in an inverted "u" shaped curve. An example of this second nonlinearity class is measures of coherence. As measured for example by the semantic similarity between sentences, we expect that incoherent (low coherence) and highly redundant (high coherence) responses will receive lower human scores implying that this type of relationship will not adequately be described by a linear model. A final example of nonlinearity is that many features reach asymptotic values, where for example appropriate use of advanced punctuation improves the quality of writing up to a point, but its contribution levels out or may even decline after that point.

The following plots demonstrate some of these issues. The distribution of student responses in feature space are represented by a kernel density smoothing of the points, where the low to high density regions of responses range from white through darkening shades of blue, with the most dense regions nearly black. The linear regression line is shown in green and a locally weighted regression curve (Loader, 1999; 2012) is shown in red.

The first plot shows how human score varies by sentence length measured as words per sentence (WPS). The density plot clearly shows that most scores fall in an area bounded between about 15 and 22 words per sentence with scores ranging from .25 to about .75 for this item. For reference, the median response is 20.2 WPS, with a lower quartile of 17.3 and an upper quartile of 24.0. We also see that WPS is not a particularly informative feature as indicated by the shallow slope of the regression line. The red, locally weighted regression line indicates that for a substantial portion of the WPS range, a linear approximation is a reasonable approximation. However, at the low end of WPS the linear model does not match human scoring, in that it awards too high scores and at the upper end of WPS, it again overestimates the contribution of WPS to human score. We see from the locally weighted regression curve that for WPS above about 25, there is no additional information on score from WPS, and in fact there may be a slight decrease in human scores at very high WPS, though this may be an artifact. This is an example of a feature reaching an asymptote described above. While there is evidence that WPS is part of the overall evaluation humans apply to the responses, it is also clear that using WPS as a feature requires additional checking to ensure the sentences are constructed in a sensible fashion.

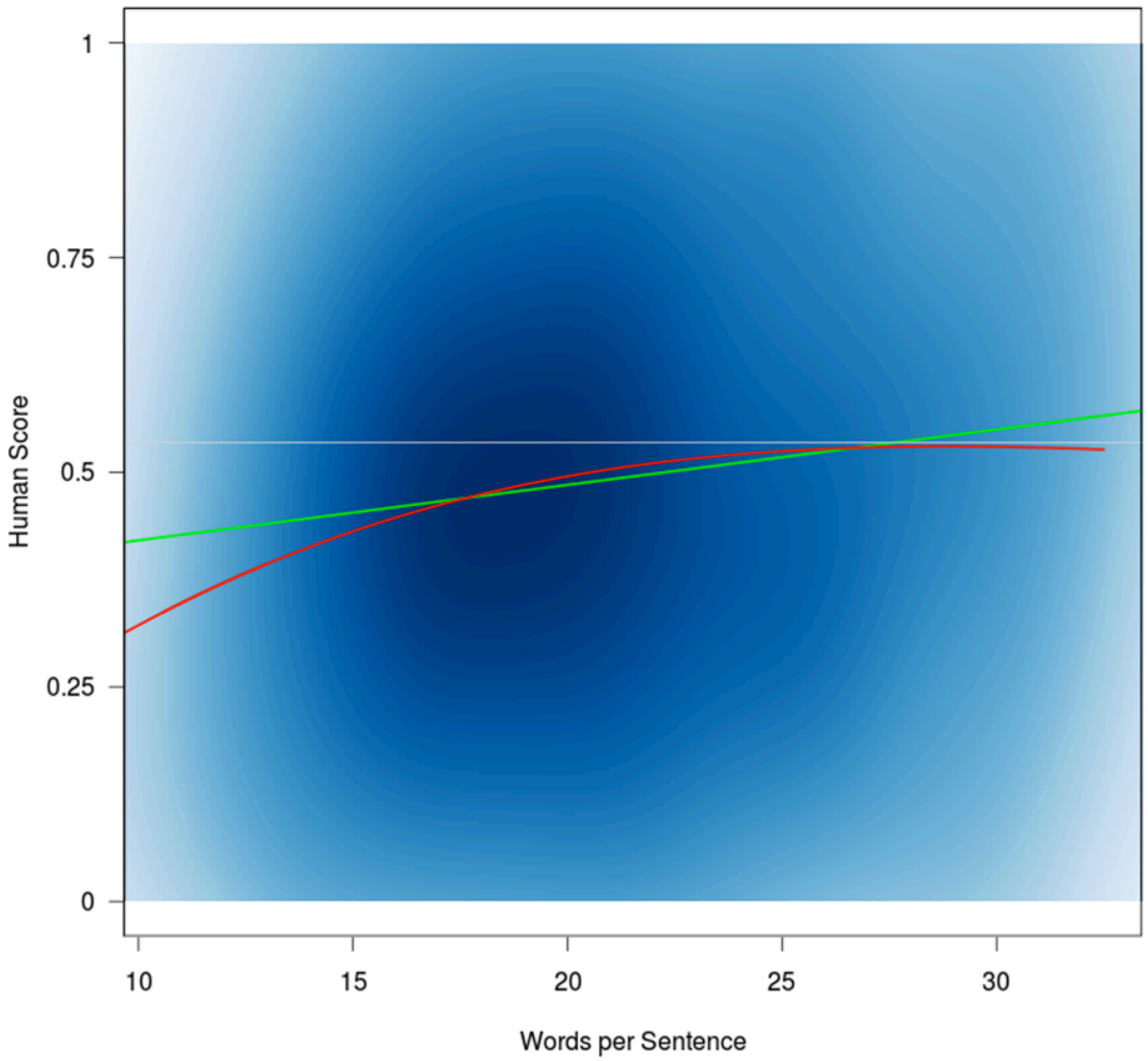

Figure 5: Smooth plot of Human Score vs. Words per Sentence. Density of responses range from white, low density, to dark blue, high density. Green line is the linear regression fit, and red curve is the locally weighted regression fit. The grey horizontal line is a visual aid to indicate the flattening of the locally weighted regression fit.

The second plot shows how human score varies with Sentence to Sentence coherence (e.g., Foltz et al., 1998). The linear model indicates a steadily decreasing score with increasing coherence, but the locally weighted regression tells a quite different story. Here for responses exhibiting low coherence, there is a region where increasing coherence is recognized and rewarded by the human scorers up to a point and from that point onward a decreasing linear model provides a reasonable approximation. An additional deviation can be seen that above coherence of about 0.5, the humans awarded lower scores at a faster rate than a linear model would allow. Again, in the case of this coherence feature, there is a large region where the linear model is a reasonable approximation, but at the low and high ends the model diverges from the evaluation of human scorers.

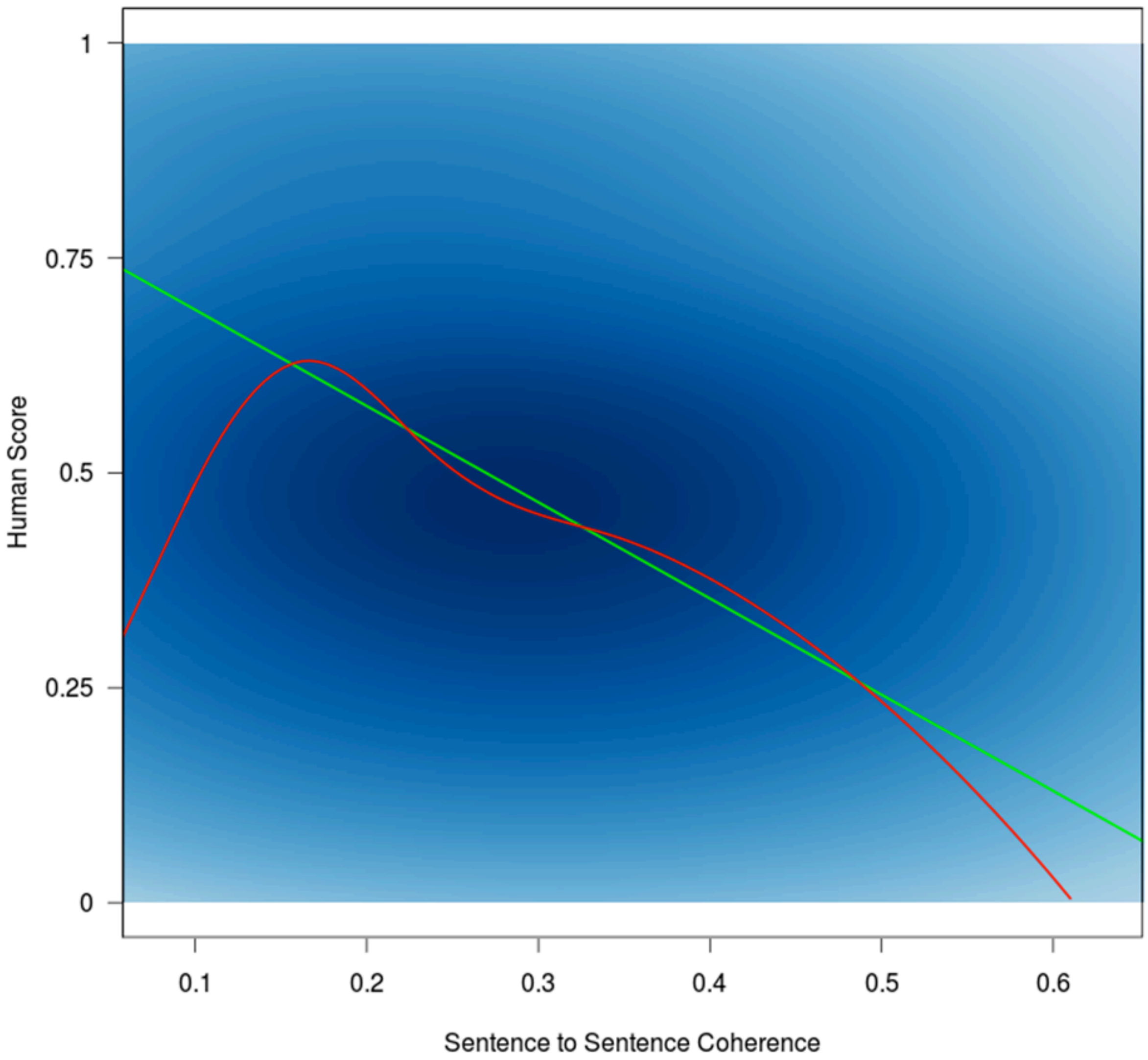

Figure 6: Smooth plot of Human Score vs. Sentence to Sentence coherence. Density of responses range from white, low density, to dark blue, high density. Green line is the linear regression fit, and red curve is the locally weighted regression fit. In the mid-range a linear model provides a reasonable approximation, but there is divergence to the human ratings at the ends.

As these two examples show, assuming a linear model could go astray, especially at feature values nearing the extremes seen in the training set. Possible solutions are using different models at the extremes or using modeling techniques that support more general functional forms.

Explanations of performance beyond length

It is well known that the length of essays for a given item often correlate highly with human scores. For instance in the ASAP study (Shermis & Hammer, 2012) of the nine sets of human scores for eight items, two sets of human scores correlated with word count above 0.8 and all correlations were above 0.5. There are sensible reasons why length serves as a proxy for attributes of the quality of a response, such as adequate content coverage requires a sufficient length to cover the topic and students without much knowledge on a topic or with low language ability typically can not generate sufficient words during a timed essay test. However, automated scoring best practice requires using more construct relevant approaches to predicting score with features less obviously tied to length than unadorned measures such as word count. In addition, overall length is a quite easy parameter to pad in bad-faith attempts.

A difficulty in implementing a policy based on downgrading the importance of length in scoring is that many useful features are highly correlated with length. For instance, the ratio of word types to word tokens gives a measure of diversity of vocabulary. However, at least for shorter essays it is also quite coupled to length. Reflecting on the mechanism of this ratio reveals that a one-word essay attains the maximum ratio of one, which initially can only decrease as the length of the essay increases, until it normalizes to provide useful information.

For an example case, we examine a class of semantic variables that are based on distances in the semantic space (e.g., Landauer et al., 2001). Distances in semantic space provide measures of the degree to which a target essay has similar content to essays from the training set. Due to the nature of the distance measure, it partially confounds with response length. For example, essays that all have similar content all tend to be of similar length. Using multidimensional scaling, we were able to generate new features based on these distance that are significantly less correlated to response length, but still allow semantic similarity to explain much of the human score.

In the following pair of plots, the actual points are identical, just the coloring is different, with responses arrayed on two new derived dimensions. The responses are colored to indicate human score, going from saturated cyan for the low scores to saturated magenta at the high end. We see that scores load quite nicely on the y-dimension. Notice the cluster of low scores at the top left. A hypothesis that is quickly validated in the right plot is that these are all very short, low-scoring essays. The right plot colors the deciles of length by word count. We see that the first component heavily loads on length. Overall this is the kind of result that indicates that it is possible to separate length from the semantic component while preserving the ability of a derived feature to predict human score based on content. Thus although essay length may insidiously influence many common features, it can be partialed out in a manner to allow measurements that are not influenced by the padding of extra words.

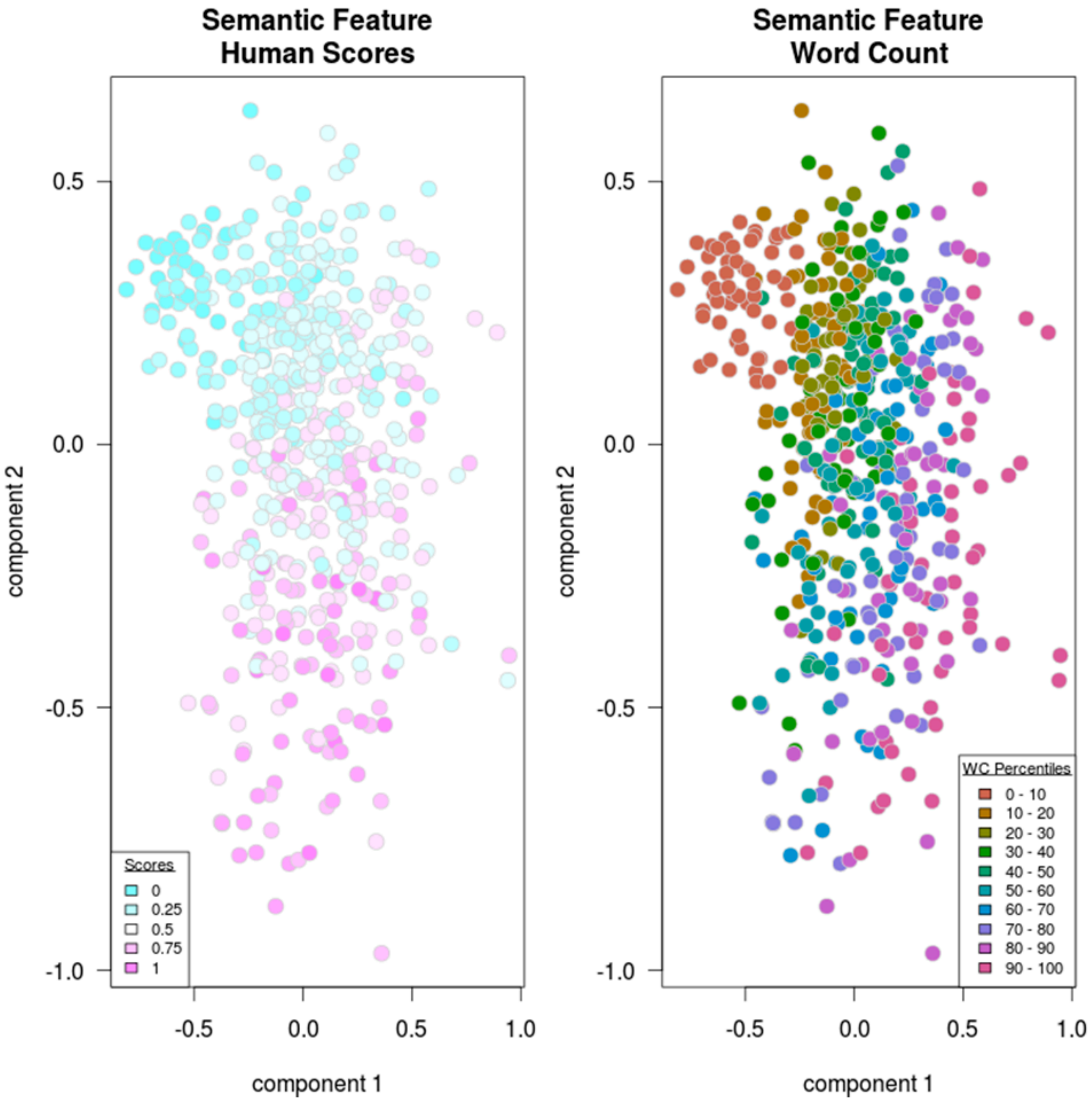

Figure 7: Responses arrayed in two derived distance dimensions. Left plot colors responses by human score, while right plot colors by length. The derived dimensions predominately separate out score and length.

Conclusions

A typical approach used in the machine learning literature (e.g. Hastie et al., 2009) is supervised learning, where a training set with a set of known response variables is modeled using a set of explanatory variables. The development of automated scoring systems has followed this strategy by collecting a number of relevant variables based on a training set of essays and then applying one or more algorithms that best predict a response variable such as student grade. In the case of automated scoring, there is a wide range of different kinds of features that can be extracted from essays and there are a number of different algorithms that can be applied to combine the variables to predict an essay score. However, the results from this paper illustrate the critical importance of understanding the assumptions that underlie the set of training data, the feature variables and the algorithms, and the need to flag essays that have characteristics that appear to violate those assumptions.

Multivariate normality provides a means to detect how well a set of features follow a normal distribution based on a set of training essays. Given the general assumption that training set essays are a representative sample of acceptable student input, confidence intervals can be set to measure the deviation of any student's input from the training set. Multivariate normality also illustrates that, although a number of machine learning algorithms (e.g., regression based approaches) tend to assume normally distributed variables, a variety of features do not follow normal distributions. This highlights the need to verify the behavior of features before applying algorithms.

An analysis of non-linearities in essay features further illustrates the need to consider the appropriateness of matching features to scoring algorithms. Features such as words per sentence will often increase from low scoring essays to mid scoring essays. However, they tend to asymptote at a certain level. Similarly, features like coherence do not closely follow a linear function. While both variables can be partially approximated by a linear model, some of richness of their contribution is lost due to their non-linearity. In addition, by actively exploiting the non-linear relationships, it becomes more viable to use them as the basis to detect outliers. For example, a measure of extremely high coherence would tend to indicate large amounts of repetition within an essay rather than high performance.

Because length can easily manipulated in trying to "game" an essay grader, it is important to investigate the effect of length on other variables and attempt to separate effects of length from the components of interest in features. We introduced an approach to controlling how the effects of the length of essays are inherently confounded with other feature variables through multidimensional scaling. The results indicated that content scoring measures can be used that operate sufficiently independently of the length of the essays.

Overall the three approaches described serve two separate functions that are critical components of a systematic effort to improve the quality of automated essay scoring. The first function is to gain a better understanding of how feature variables and algorithms inter-operate in order to ensure that they are used within the range of assumptions that ensure accurate scoring. The second function is use the information from assumptions of the feature variables and algorithms to be able to detect when a set of features in an essay violate the assumptions of the model. This violation could be due to gaming, inappropriate input, or just new essays that are significantly different from the training set. In all of these case though, the approaches can contribute to the development of validations that flag essays as being beyond the bounds of what the model was intended to score.

References

Cox, D. R., and Small, N. J. H. (1978). Testing multivariate normality. *Biometrika* 65 (2): 263.

Foltz, P. W., Lochbaum, K.E., Rosenstein, M. B., Davis, L. (2012) Increasing Reliability Throughout the Automated Scoring Development Process. Paper presented at the *National Council for Measurement in Education Annual Conference*, Vancouver, CA, April.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25*(2&3), 285-307.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning (2nd ed.)*. New York:Springer.

Healy, M. (1968). Multivariate normal plotting. Journal of the Royal Statistical Society, Series C. Vol 17. No. 2. pp 157-161.

Jurafsky, D. and Martin, J.H. (2009). Speech and Language Processing. Upper Saddle River, NJ: Pearson.

Landauer, T. K., Laham, D. & Foltz, P. W. (2001). Automated essay scoring. *IEEE Intelligent Systems.* September/October.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer.

Loader, C. (2012). locfit: Local Regression, Likelihood and Density Estimation. R package version 1.5-8. http://CRAN.R-project.org/package=locfit.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* 2 (1): 49–55.

R Core Team (2012). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shermis, M. and Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Paper presented at *NCME annual meeting*, Vancouver, CA.

Williamson, D. M., Bennett, R., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D., Way, D., and Sweeney, K. (2010, June). *Automated Scoring for the Assessment of Common Core Standards.*