Automated Essay Scoring Using Machine Learning

Shihui Song SHIHUI@STANFORD.EDU
Jason Zhao JLZHAO@STANFORD.EDU

Abstract

We built an automated essay scoring system to score approximately 13,000 essay from an online Machine Learning competition on Kaggle.com. There are 8 different essay topics and as such, the essays were divided into 8 sets which differed significantly in their responses to the our features and evaluation. Our focus for this essay grading was the style of the essay, to which we extended by adding the category of maturity. We evaluated Linear Regression, Regression Tree, Linear Discriminant Analysis, and Support Vector Machines on our features and discovered that SVM achieved the best results with an average $\kappa = 0.78$.

1. Introduction

The automated essay scoring model is a topic of interest in both linguistics and Machine Learning. The model systematically classifies the quality of writing and can be applied in both academia and large industrial organizations to improve operational efficiency.

1.1. Motivation

Each year, thousands of students take standardized tests with the same essay topics. Hand grading these essays is tedious and subjective. Instead, many organizations have already turned to automated essay grading to improve consistency and efficiency. Accurate models will not only reduce the amount of human error/variance in essay grading but could also save school boards and teachers many precious hours that could be used to improve the educational system.

CS224N Final Project, *Shihui Song, Jason Zhao*

2. Data Set

The training and test data were acquired from a past competition from Kaggle.com [1] sponsored by Hewitt-Packard. We had approximately 13,000 number of essays ranging from 150-550 words each provided for us. We split the essays into a 70-30 training and validation scheme, which results in a size of 9,100 essays for the training set and 3,900 essay for the test set. This divides further into around 1,200 training essays and 500 test sets per essay set.

3. Feature Generation

Python was utilized for the pre-processing of the data into matrices that were then fed to Matlab for supervised learning. The Parts of Speech Tagging from the Natural Language Toolkit (NLTK) was the sole library used for the assignment. Most of these style categories are based of a recent ACL paper classifying goodness of scientific New York Times articles (Louis & Nenkova, 2013).

There were five categories of features that were considered and generated for this project for style.

3.1. Its visual nature

A descriptive sentence awes the reader and prompts their imagination. The source of imagery and meaningfulness used for this data set is derived from the British Natural Corpus where each word has a imagery score between 0 and 999 (Kilgraff, 1996). The features derived from this set include the proportion of words that are visual, proportion of unique visual words, average imagery scores for those words, the average imagery score for the essay, and all of the above for every third of the essay by dividing the essay into an introduction, a middle, and a conclusion.

3.2. Inclusion of pronouns

Just as scientific articles which explicitly reference people or experiments would most likely be more re-

[1] http://www.kaggle.com/c/asap-aes

spectable and concrete, we imagined that similar essays in our data set with external reference would score better. The Kaggle essays already contained name entity tags from the Name Entity Recognizer from the Stanford NLP group. Name entities for PERSON, ORGANIZATION, and LOCATION were categorized as *proper pronouns*, there were also counts of *personal pronouns* such as "he", "myself", etc, and the last pronoun count was for *relative pronouns*, which were noun phrases (tagged by the python NLTK tagger) followed by "who", "which", and "where".

We then divided all people pronouns into *animate* and organizations and location into *inanimate* under the assumption that the use of people would be more engaging than locations and organizations.

3.3. Its beautiful words

Beautiful word choices are thought to increase an essay's elegance, thereby its score. Only words above 5 characters in length are considered beautiful for this project. Two factors are considered for individual word beauty:

- **High perplexity-letter model** - how unlikely the combination of characters in the word is. For a word, find the product of its character frequencies (Cornell Math Cryptography, 2003). The lower the product the more complex the word.
- **High perplexity-phoneme model** - we created a 4-gram for syllables to determine phoneme frequency (CMU, 1998). For every word in the essay, check to see if there exists a pronunciation for it, and if so, then also find the likeliness of its 4-gram combination.

The ultimate features for this category were the average letter and phoneme frequencies per beautiful word and per essay, as well as the top 10, 20, and 30 average phoneme and letter frequencies where the top is the lowest frequency.

3.4. Its emotive effectiveness

A very dry and emotionless essay is not powerful. The Subjectivity Lexicon from MPQA provides a list of words and their sentiments (positive, negative, neutral, or both) and the strength of those sentiments (MPQA, 2005). The resulting features are proportions of sentiments and strength individually and combined for the entire essay or for given sentiments and strength. In addition, we also calculated the proportions of different emotions to one another.

In the end however, we realized that exhausting every combination of sentiment and strength proportional to another was actually noise that hurt the Kappa score. Therefore, we removed all proportions of sentiment and strength to other sentiments and strengths.

3.5. Its maturity

Our vocabulary expands as we grow older. Therefore, in a sense, as we mature, so does our vocabulary. This is particularly poignant for this set of essay, as they are written by adolescent students.

Although this isn't included "style", we thought this would be a good additional category. The Age of Acquisition is the average age when a person learns the certain word (Kuperman, 2012). For every essay, we found its average maturity, top mature tokens, and vocabulary maturity.

4. Learning Algorithms

For the project, we evaluated several different classes of learning algorithms which will be described below. Most of the algorithms we evaluated are regression based where we treat the essay scores as a range of values and predict a floating point value within that range.

4.1. Linear Regression

We used a simple linear regression model implemented by the statistics packet in Matlab (The MathWorks, 2013). The LinearModel class fits a linear function $h_\theta(x) = \theta^T x + c$ to a design matrix X in order to minimize the least square error as discussed in class.

4.2. SVM

We used the ν-SVM algorithm presented by (Schölkopf et al., 2000). It is a regression SVM algorithm based on the ϵ-SVM which introduces the ξ_i slack variables for capturing error (Vapnik, 1995). Specifically, the ν-SVM attempts to solve the following problem:

$$\min \tau(\mathbf{w}, \xi^{(*)}, \epsilon) = \frac{1}{2}||\mathbf{w}||^2 + C\left(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right)$$

$$s.t.\ ((\mathbf{w} \times x_i) + b) - y_i \leq \epsilon - \xi_i$$
$$y_i - ((\mathbf{w} \times x_i) + b) \leq \epsilon + \xi_i^*$$
$$\xi_i^{(*)} \geq 0, \epsilon \geq 0$$

We chose the ν-SVM because it reparameterizes the loss sensitivity term ϵ in the tradition C-SVM. This is desirable because ϵ is very hard to tune in practice whereas ν is simply an upper bound between the

training error and the number of support vectors.

4.3. Multiclass Linear Discriminate Analysis

We used Matlab's ClassificationDiscriminant class to train a multi-class linear discriminant classifier(The MathWorks, 2013). The classifier predicts new examples using the following rule:

$$\hat{y} = arg \min_{y=1,...K} \sum_{k=1}^{K} \hat{P}(k|x)C(y|k)$$

where K is the number of classes, $\hat{P}(k|x)$ is the posterior probability of k given x and $C(y|k)$ is the cost of classifying y when true class is k. We choose to use the default cost matrix for the milestone but may investigate other cost matrices for the final report. The posterior probability is estimated using a multivariate Gaussian distribution where the mean μ_k and covariance matrix Σ_k are approximated from the training set.

$$P(x|k) = \frac{1}{2\pi|\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$$

4.4. Regression Trees

Matlab also supports a binary splitting decision tree that can fit to some response variable Y. This is an interesting algorithm because it uses a recursive partitioning model to divide the feature space into simple buckets. We used the mean squared error as the splitting criterion with a minimum leaf size of a single observation. The tree is post-pruned to generate the optimal sequence of subtrees. The resulting geometric interpretation is that the feature space is split into linear boxes (since this is a binary regression tree). Therefore the end result similar to an unsupervised clustering algorithm but the training phase is drastically different.

5. Experimental Results and Algorithm Selection

5.1. Error Rates

The measure of error rate utilized is the quadratic weighted Kappa. The quadratic weighted Kappa measures the agreement between the automated essay grader and the human scores. Scores typically range between 0 (random agreement) and 1 (total agreement), although scores less than 0 can occur when there's less agreement than random.

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j}O_{i,j}}{\sum_{i,j} w_{i,j}E_{i,j}}$$

The weight is the squared difference of the i and j scores over the square of the size of the set minus one. E is just a calculation based on the user's given vector based on score frequency. And $O_{i,j}$ is the number of occurrences when score i is assigned by grader one and score j is assigned by grader 2.(Kaggle.com, 2012)

5.2. Algorithm Performance

We evaluated the different algorithms using essay set 1. For the initial feature vector, we used the words that appear more than 5 times across all essays because certain algorithms such as linear regression and linear discriminate analysis cannot handle matrices with 40000 columns. For plotting the learning curve, we measured the Kappa error as we increased the training set size.

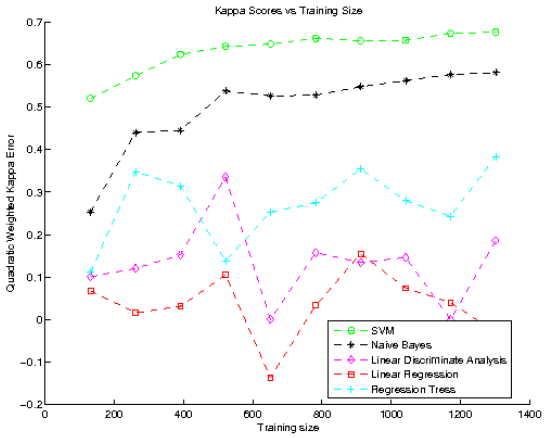

Figure 1. The Kappa error for each different learning algorithm over set 1.

As shown in figure 1, SVM performed the best over essay set 1 with Naive Bayes following in second. For these 2 algorithms, it seems training error decreases as training set size increases which is desirable. For all of our subsequence evaluations, we decided to focus only on the SVM.

6. Evaluation and Analysis

Below is the Kappa score comparison for the feature categories of style and their scores for each essay set. Please note that essay set 2 and 8 are excluded because essay set 2 asked for two separate ratings, which were furthermore dependent on specific trait rubrics. Essay set 8 was also not included due to the Matlab machine learning algorithms becoming rank deficient during the process and we felt that it was not a whole-

some measurement.

6.1. Individual features

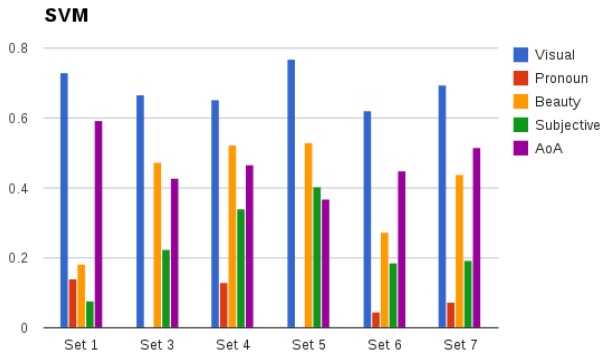

Figure 2. Individual style category and their Kappa scores across each essay set

Of the various individual feature sets in the graphs above, imagery has consistently the highest kappa score. Beauty and maturity are both respectable while subjective is dismal. The worse is pronouns, which does as worse as random at time. These results were not on par with our original assumptions.

Pronoun was a complete disappointment, leading to our conclusion that in these adolescent essays, references to entities does not matter.

We did not have too much hope for beauty, as the phonemes and character frequency collected by us were not thought to be as great statistically as those given by the other feature categories; however, we were pleasantly surprised with the results. Upon closer analysis, we realized that often, beauty scores are often synonymous to word frequency scores, which we had already proven to be a good metric.

Subjectivity scores depends entirely on the prompt. Essay set 5 asks the reader to describe how the setting affects the character, thereby asking for emotions. When we delved into the results further, in many of the persuasive essay sets, negative essays scored better in general than positive ones. This leads us to believe that criticism is more highly regarded than praise.

The maturity scores (AoA) were also delighted to have maturity be such a good indicator. The prompts with more freedom such as narrative essays in essay sets 1 and 7 achieved higher Kappa scores with maturity. This could very well be caused by the range that students are allowed to express themselves in, not restricted by a necessary rubric to follow.

When reviewing the categories as a whole, we discovered that the prompt of the essay has a huge affect on the score of the categories. Essays that need to address a certain topic or make a point, does not need to have style to score high. Other prompts with freedom of expression often has style factored in to the final score. Thus, our style score depended on the prompts.

6.2. Algorithm Tuning

In order to achieve the lowest training error, we performed a 10-fold cross validation over essay set 1 in order to find the best values for C and ν. The plots generated below shows the effect of these 2 variables for the Kappa score.

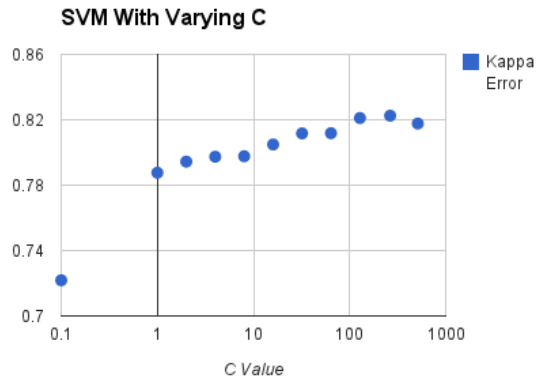

Figure 3. Effect of the parameter C on SVM prediction accuracy.

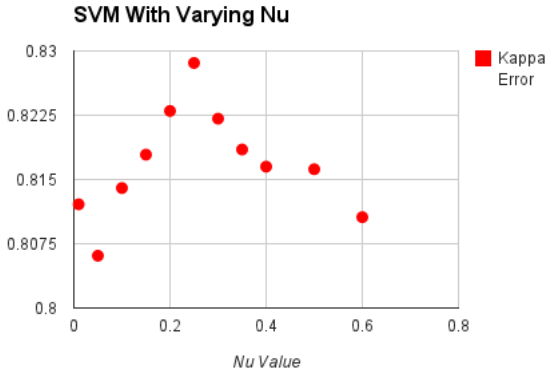

Figure 4. Effect of the parameter nu on SVM prediction accuracy.

This shows that we perform better with a C value of around 256 and a ν value of around 0.25. This means we penalize incorrect entries greatly and would like to approximate a hard margin SVM as closely as possible. We also experimented with various Kernel functions and found that a Radial Basis Function performed the best for essay prediction.

6.3. Final Result

We used our optimal training parameters to classify the essays in each of the sets. Our results are shown in the table below.

Essay Set	Kappa Score
1	0.8286
3	0.6511
4	0.734
5	0.7789
6	0.6934
7	0.6762

7. Conclusion and Future Work

While our project has had certain success in this domain of automated essay grading for adolescents, it is not without pitfalls. Judging by style itself is not a good indication of the paper addressing the prompt. To truly judge an essay we must understand at an artificially intelligent level both the prompt and the essay.

However, this project takes strides in what we believe to be understanding the style of adolescents. Our inclusion of maturity has proved to be a dependable category feature to judge by, an addition that could also perform very well in any other essay grading.

This project has also followed closely on the features of the style according to the Louis and Nenkova paper, but perhaps in the future, more explorations of what composes of "style" would be a good venture. As an extension, it would be worthwhile to explore not only the style by token content (as this project has mainly done so so far), but also on the structural and syntactical style.

References

Automated essay grading using machine learning. 2012. URL http://cs229.stanford.edu/proj2012/MahanaJohnsApte-AutomatedEssayGradingUsingMachineLearning.pdf.

CMU. Cmu pronounciation dictionary, 1998. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Cornell Math Cryptography, Group. English letter frequency, 2003. URL http://www.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html.

Kaggle.com. The hewlett foundation: Atumated essay scoring - evaluation, February 2012. URL http://www.kaggle.com/c/asap-aes/details/evaluation.

Kilgraff, Adam. Bnc database and word frequency lists, 1996. URL http://www.kilgarriff.co.uk/bnc-readme.html.

Kuperman, V. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 2012. URL http://link.springer.com/article/10.3758%2Fs13428-012-0210-4/fulltext.html.

Louis, A. and Nenkova, A. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association of Computational Linguistics*, 2013. URL paperhttp://www.transacl.org/wp-content/uploads/2013/07/paper341.pdf.

MPQA. Subjectivity lexicon, 2005. URL http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.

Schölkopf, Bernhard, Smola, Alex J., Williamson, Robert C., and Bartlett, Peter L. New support vector algorithms. *Neural Comput.*, 12(5):1207–1245, May 2000. ISSN 0899-7667. doi: 10.1162/089976600300015565. URL http://dx.doi.org/10.1162/089976600300015565.

The MathWorks, Inc. Supervised learning, 2013. URL http://www.mathworks.com/help/stats/supervised-learning.html.

Vapnik, Vladimir N. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.