Exploring Potential for Machine Learning on Dataset about K-12 Teacher Professional Development

Devney Hamilton Keziah Plattner

December 14, 2013

1 Problem and Goals:

Evaluation of k-12 teachers in the United States involves intense political and philosophical debates. Data on teacher professional development (PD) practices is only just emerging in digital form, as PD management for k-12 moves to online platforms, and teachers are rated on numerically-scaled rubrics. Education leaders hope to identify generalizable factors in PD practices, previous experience, training, and institutional support that predict teachers' improvement. But first we mustdetermine the viability of currently available data for predicting teachers' ratings.

We began this project with a dataset recording teachers' professional development activities and ratings of their performance in one large charter school network, drawn from a commercial PD-management platform. Our goal was to find which available features are relevant to predicting ratings, identify barriers to meaningful analysis of the dataset, and make recommendations to interested education organizations on data collection for future analysis.

2 Setup

2.1 Data and Y-Values

Some data points were associated with a teacher, some with a coach, and some with an observation event. We decided to define a sample as a teacher because there were more teachers than coaches, and more features associated with teachers. Each teacher is rated on a scale of 1-4 for some subset of 30 rubric items, each of which describe one aspect of teaching. We assigned each teacher a single y-value that was equal to the average of these indicators.

Our samples were clustered tightly around the mean score of 2.9, so we converted the continuous y values into two classes with class 0 being 'below average' (score less than 2.9) and class 1 being 'above average' (score greater than or equal to 3.0)[1]. We threw out the 15 samples that fell directly on the average.

2.2 Features

We used subsets of these per-teacher features throughout our experiments:
A) volume of evidence in their evaluations
B) number of professional development activities they have engaged in
C) the number of teachers their coach works with
D) number of rubric items they were rated on
E) number of resources teacher attempted
F) completed resources per teacher
G) goals completed
H) number of goals made
I) the average length in days of all their observations
J) what month of year they were observed
K) which week of the year they were observed

2.3 Cross-Validation Strategy

We tried several different strategies with training on our data. First, we did random cross validation holding out 20% of the data, then iterated over many runs of this to find the average error on our training and testing data. Next, we tried leave one out, and got very similar error rates. Ultimately, we

[1]We used this split with all our experiments unless otherwise specified

decided on one iteration of cross validation holding 30% of our data, so we could analyze the training and testing data in order to better understand our error rates.[2]

3 Initial Attempts

We initially started experiments with only features A-E. We tested linear regression with the raw average scores, but quickly found out that the data was not linear.

SVM and Naive Bayes (NB) were readily available in MATLAB, and were our next starting point. Given that most of our data involved counts of tokens (such as number of goals, resources, etc), a multinomial distribution made the most sense. In addition, the multinomial distribution makes an assumption that the total number of tokens is independent of the response class. This fit the best with our data, since each teacher has a different amount of information about them, but it should not impact how they are scored in the end. With this logic, a sample is a teacher's collection of PD actions. The prior on the data set as a whole is 40% class 0 and 60% class 1, and our initial experiments with SVM[3] and Multinomial NB (MNNB) left us with a minimum development error of 34%.

This low performance left us to explore in our further experiments whether it was due to an insufficient number of samples, bad feature selection, or a bad modeling of the problem. In these experiments we faced a constant trade-off between more complete information per-sample and large sample size.

4 Further Experiments

4.1 Additional features

Our next step was to exhaust our database of features, resulting in a total of 11 features. Unfortunately, some information like teachers' years experience, communication among teachers and coaches, and the specificity of teachers' goals were not available as we hoped.

In addition, many of the additional features were present for only some of the teachers. We replaced the counts of missing features with zero[4] in order to maintain our larger sample size. Our baseline results for features A - K using imputation of missing data was:

Dataset:	Train	Dev
MNNB error	.33	.36
SVM error	.31	.38
Class 1 Prior	.59	.62
Sample Size	287	122

	Class 0	Class 1	% diff	n
Count of collected evidence	4.07E-02	5.36E-02	32%	509
Count of p.d. Activities teacher did	2.07E-02	2.94E-02	42%	424
count of observations coach responsible for	1.37E-01	1.94E-01	42%	125
Count of rubric items evaluated	1.10E-01	1.00E-01	-9%	424
Count of pd resources used	3.24E-02	5.33E-02	65%	295
Count of pd resources completed	1.02E-02	1.11E-02	9%	114
Count of goals completed	1.40E-03	2.15E-03	54%	27
Count of goals created	3.45E-02	4.84E-02	40%	370
Duration in days of observation	3.25E-01	2.11E-01	-35%	415
Months of year before observation	6.52E-02	6.70E-02	3%	415

Figure 1: Features with weights most different between classes are highlighted in blue. Most are available for only a subset of the data. The relate to the teachers' and the their coaches' professionalism and involvement in PD activities. Most have values so small they are likely to be overridden by the prior.

The weights are reported in Figure 1, where n is the number of samples for which a feature value was available.

Only a small subset of teachers had features like the number of completed goals, and the number of times they had used professional development resources. However, these features had the greatest distinction between above average teachers and below average teachers. For example, the number of

[2]The reported results in this paper come from this last strategy.

[3]Unless otherwise noted, SVM was always run with a linear kernel, as our experiments with other kernels provided no improvement.

[4]See Interpolation Experiments for more details on our attempts with generative classifiers and imputation.

goals completed was 54% higher for teachers who were above average, the number of goals created was 40% higher, and the number of resources completed was 65% higher. Unfortunately, this data was rarely available for the samples we were predicting, resulting in a higher error rate in the development data set compared to the training dataset.

These significant differences in weights despite our high error rate implies that these are key features in determining teacher performance, but are either only a subset of important features needed to more accurately predict teacher performance, or need to be present in all our data in order to accurately predict teacher performance on testing data.

4.2 Try redefining y-values

Teachers are evaluated on rubrics. Our initial attempts ranked teachers on the average of all their scores. However, the rubrics are broken into 4 specific categories. We tried defining a teacher's label based on their score in each category. We ended up getting better results with the category labeled 'Professionalism', averaging around 27% development error with MNNB and 25% with SVM (102 samples, 70% prior on class 1).

This makes sense for two reasons. Features that show a difference between below- and above-average teachers are related to professionalism practices (goal setting, resource usage). Another explanation is the 30-70 prior split, making this performance little better than what we observed with our more general definition of a y-value.

4.3 Limiting the dataset

We were curious if we could get better results on the subset of data for which we had values for features that had distinguishable weight values. Requiring all samples to have completed goals and completed resources would have restricted our datasize to so small that randomness would have overridden results. Instead, we used the subset (n = 200) for which simply the count of goals and count of resources were available:

Dataset:	Train	Dev
MNNB error	.45	.40
SVM error	.38	.50
Class 1 Prior	.57	.48
Sample Size	140	60

The influence of feature weights in the larger set decreased so that all the weights for class 0 and class 1 were nearly equal. This suggests that the presence of those additional features, while incomplete, is significant in classifying teacher performance, even if we must fill in missing data for the rest of the samples. Our following imputation experiment suggests the absence of some data is itself meaningful.

4.4 Imputation

We had choices about how to handle missing feature values. Because our features were counts from a MySQL database where the count starts as soon as something is created (a goal, an activity, etc), we set NULL values as '0' at first. This is consistent in what we observed in 4.3, where the absence of goal and resource data in samples mattered more than variations within the subset where that data was available.

We wanted to test if this 0-count assumption was breaking existing correlations in the features, so we also tried using imputation by filling missing values with the average for that feature. Unfortunately, we did not find a performance improvement over replacing missing values with 0, comparing to our baseline in 4.1. This is consistent with our assumption that our features are inherently counts of actions.

Dataset:	Train	Dev
MNNB error	.45	.35
SVM error	.40	.46
Class 1 Prior	.49	.46
Sample Size	221	94

We also experimented with latent variables by adding a boolean for whether or not features were available. In this initial attempt we saw no noticible difference, but this needs further investigation.

5 Model Fit

In our experiments with more features and different sample sizes, we tried SVM with various kernels, but were never able to bring its performance above that of MNNB. We used this as a suggestion that our choice of MNNB as a model was not a limiting factor in our performance.

However, we later tried classifying only extreme examples, defining class 0 as an average score of less than 2.5, and class 1 as an average score of greater than 3.5. We replaced missing feature values with the average of that feature over included samples. We were were surprised to find that it was easier for SVM to classify a subset of this dataset where the labels are extremely different, though SVM's performance on the full dataset had been indistinguishable from that of MNNB. Our results are summarized below:

Dataset:	Train	Dev
MNNB error	.33	.50
SVM error	.222	.366
Class 1 Prior	.333	.116
Sample Size	72	30

The low performance in MNNB may be in part due to the small sample size imposed by the restrictive definition of the classes. On the training set, MNNB performed equivalently to assigning all samples class 0, since the error is equal to the prior on Class 1 (.33). On the development test set, it performed no better than a coin flip. For the training set, SVM performed better than the prior. With the development test set, SVM also performed better than MNNB.

SVM's superior performance suggests that with classifying extreme examples, we can find a model that is better than than the multinomial model, since in general SVM classifies with fewer assumptions than MNNB. The dataset with extreme examples is more likely to be separable than the dataset including samples close to the mean (see over-all analysis). Since SVM did better than MNNB in this case, and better than MNNB ever did on a fuller set including samples closer to the mean, it suggests that the extreme examples are better separated with an SVM model than with an MNNB model. In this case of more-likely-separable data, it seems that perhaps the MNNB assumptions did pose a barrier to performance when compared SVM.

This experiment also suggests that it would be useful to model latent variables that might help distinguish samples whose labels fall closer to the mean, and expose variance in teachers' intrinsic performance level.

6 Over-all Analysis

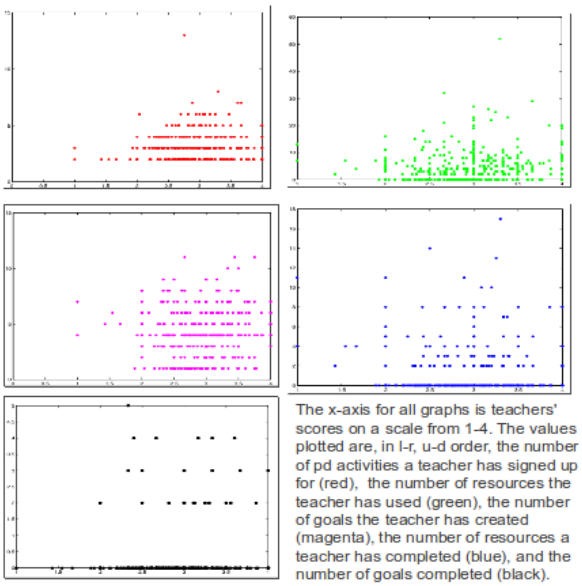

Figure 2: Plotting features that have distinguishable weights in our baseline (4.1) and have large sample sizes creates a pattern filling the lower right half of the plot. This helps us understand why were were more successful classifying only extreme examples, and the need to find other variables (latent or observed in other domains) affecting teachers' ratings.

We had somewhat more success classifying extreme examples. MNNB found differences in the counts of certain PD-related activities for below and above-average teachers. However, these differences were not enough for successful classification. We decided to plot these more interesting features against teachers' scores (average over their rubric ratings, before classifying into Class 0 and Class 1). What we found is that more PD activities predicts a higher score. However, a high score does not predict more PD activities, as demonstrated in Figure 2. In this figure, we see that three of the five most influential weights (difference between classes is more then 40% in our baseline) in our baseline fill the lower right triangle of the plot: the number of PD activities (red), the number of goals (magenta), the number of resources used(green). The number of resources completed(blue) also falls roughly in

the lower-right triangle, but was not a feature that got noticably different weights, perhaps because of the relatively low number of samples for which that information is available.

This supports our suspicion that there are other factors that have an important influence on teachers' scores. These may be latent in PD data, and discoverable via more complex models, or they may be observable in other domains such as teachers' experience, training, and institutional support. It is also consistent with our finding that it is easier to classify more extreme examples, since very low scoring teachers generally have taken very few PD actions.

7 Conclusion

With the currently available data and preliminary experiments, we were able to identify which features were more and less relevant to teachers' ratings. We saw in 4.1 that certain features had a big difference between different classes, but our error rate was still high. This means that these features are valuable in understanding teacher performance, but are only a subset of the total features needed in order to accurately predict teacher performance. This conclusion fits with our other experiments. Extreme examples are easier to classify, since they don't need the extra nuances of other features in order to correctly predict teacher performance. Since extreme examples are more likely to be linearly separable given observable features than the dataset as a whole, there may be latent variables at work in examples closer to the mean. The missing feature data in some of the teachers is not necessarily problem, since just the absence of relevant features can help predict scores. Lastly, our experiments in redefining of our y-value indicates that the observed features we were able to work with are mostly relevant to a specific area of teacher performance, 'Professionalism'.

Overall, it was both exciting and frustrating to work with real-world, unexplored data!

7.1 Implications

Confidentiality issues in education can impede statistical analysis: when we requested more relevant features such as teachers' years experience and qualifications, the organization could not provide them. To make meaningful conclusions about teacher performance, organizations should prioritize combining relevant information for statistical analysis. This conclusion is confirmed by several education research sources.[5][6] Also, PD management providers should be aware that there is a difference between the data necessary for managing PD and that for describing PD practices. Timestamps are important for scheduling, but fortunately had little relationship with teachers' scores. Other expected features were not present. For example the platform had the potential to measure teacher-coach interaction and collect teacher metadata, but since users chose to not use that functionality, the tables for this data were nearly empty. In this case, the database optimized for personal applications rather than global queries, so there was significant per-feature overhead. Resolving these issues requires cooperation with schools to combine data sources and allowing for data processing overhead.

7.2 Next steps

The next steps to predicting teachers' ratings are to include more relevant, observable features. We also recommend further exploring latent variables with models such as a mixture of gaussians or hidden Markov models. . The distribution of scores suggests there is social pressure to rate teachers near the mean (since deviations have funding and employment implications), obscuring naturally occuring variance in teachers' effectiveness. However, the implementation overhead suggests this should wait until a richer dataset is available.

Our later experiments suggest that logistic regression or further-tuned SVM could better for classifying these samples than MVNB.

[5]National Education Policy Center.RESEARCH-BASED OPTIONS FOR EDUCATION POLICYMAKING : Teacher Evaluation. William Mathis, University of Colorado, Boulder. September 2012.

[6]Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems. Joan L. Herman, Margaret Heritage, and Pete Goldschmidt