Exploring Potential for Machine Learning on Data about k-12 Teacher Professional Development

Devney Hamilton (devney) and Keziah Plattner (keziah)

Recap of proposal

We're exploring a new, yet-untested data set on teacher professional development behavior and performance evaluations. The goals are to

- identify what features predict teachers' current evaluations
- find what dependencies and independencies are in this data set
- make some recommendation about what features and algorithms will be most useful for understanding the connection between teachers' professional development practices, their current ratings, and (when more years of data are available) improvement in their ratings.

Progress overview:

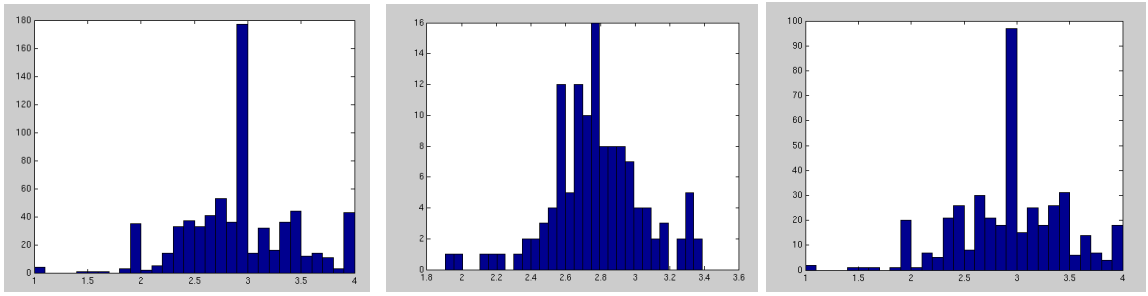
We have extracted our data from our MySQL database into samples that each represent a teacher. Each sample is labeled with an average of that teacher's ratings. We tried Naive Bayes and SVM on a limited feature set. Our errors so far are high, but data exploration suggests that we can add features and normalize features.

Data Extraction Strategy:

We received our data in a MySQL database that is highly normalized and designed for a web application for individual users. Our first attempt to adapt existing application code (in PHP) to extract features and labels failed and actually wasted a couple days of work. Our second, more successful, approach was to dust off our MySQL 'join' skills and write our own queries in MySQL Workbench, and export results to csv files.

Initial Observations of Data

As with any 'real world' and yet-unexplored dataset, many values are missing for both features and labels. We considered 2 possibilities for the definition of a sample: a teacher or an observation. Most teachers have 1 or 2 observations. We made a histogram of average ratings by teacher and by observation, and found they were almost identical, with a mean of about 2.9 and a somewhat normal distribution on either side of the mean. The histogram for average ratings across coaches is more normally distributed. We hope to investigate why this. We decided to use a 'teacher' as a sample, of which we have 424 labeled samples. We have divided these up into 60-20-20 training, development, and test sets.



Distribution of evaluation scores by observation, coach, and teacher, respectively.

Features:

So far in CSV format we have the following features for each teacher:

A) number of professional development activities they have engaged in
B)volume of evidence associated with their evaluations
C)grade levels (hopefully not correlated with performance)
D)subject taught (hopefully not correlated with performance)
E)how busy their coach is (in number of teachers coach works with)
F)number of rubric items they were rated on

Examples of features we hoped for but have found empty in database: years experience, volume of communication between coach and teacher, number and specificity of goals teachers set for themselves.

Labels:

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. Since there are only subsets for each teacher, we imputated and assigned each teacher a single y value that is equal to the average of whatever indicators they were rated for.

Initial attempts:

So far we have tested naive bayes and svm (with matlab packages) with two classes: a score of strictly below 2.9 and a score greater than or equal to 3. We omit samples closest to the mean. These our are error rates without tuning parameters:

	With only feature B	With features A, B, C	With features A, B, C omit outliers	With features A, B, C, Omit outliers and normalize
naive bayes	43%	39%	38%	42%
svm	56%	35%	37%	38%

In order to get these numbers, we did a form of cross validation by randomly selecting a new group from our data (60% of our data each time), training naive bayes/svm on it, then finding the error. We did this iteration around 100 times and found our average error.

What We've Learned So Far

After adding more features, SVM is working slightly better. This is probably because the correlation between features and y labels is very low, so samples close to the margin are more important.

Adding features improves error rate.

Basic approaches to omitting outliers and normalizing did not improve our performance.

In this data set, it is actually useful to know that certain features do *not* contribute to predicting teachers' ratings. Knowing that some features are not useful helps us understand to what extent teacher evaluation schemes are measuring teacher performance rather than noise or being correlated with aspects of the evaluation process.

Potential Next Steps

- Add in features for: time span of observations, time of year of observations, class of rubric topics used in observation (potentially relatable to education literature on areas of teaching that are most difficult).
- Plot features and y-values for correctly and incorrectly classified data to see if there is a difference.
- Try leave-one-out to see if more training data helps.
- Examine feature weights produced by NB and SVM.
- Try linear regression on continuously valued y-labels.
- Explore dividing data by rubric indicator classes and using clustering.
- Contact data source for other features such as years experience and accreditation of teachers.
- Relate to other statistical work on teacher evaluation
- Find research on classifying professional performance evaluations.