Classifying Substance Abuse among Young Teens

Dylan Rhodes, Sunet: dylanr

December 14, 2012

Abstract

This project attempts to use machine learning to classify substance abuse among young teens. It makes use of a dataset published by the Substance Abuse and Mental Health Data Archive for open use and analysis. The dataset comprises a nationally representative sample of 9227 American schoolchildren with the results of a questionnaire posing topics such as home life, eating habits, family structure, hobbies, personality/attitude, school life, bullying, and physical activity as well as the use of alcohol, tobacco, marijuana, and other controlled substances. Over the course of the project so far, three machine learning algorithms were trained, tested, and compared: logistic regression, unweighted linear regression, and a kernelized support vector machine. As a baseline, the naive classifier, which always predicts -1 (no substance abuse), was compared to the three algorithms. Ultimately, the kernelized support vector machine displayed slightly higher classification accuracy with properly tuned parameters, but logistic regression offered similar results with a lower time complexity.

1 Introduction

As previously stated, the goal of this project was to use lifestyle factors to predict substance abuse among young teens. This is an important task for several, fairly obvious reasons. First, early substance use, besides being objectively undesirable in and of itself, has been shown to indicate problems later in life including delinquent behavior and later abuse.[1, 4] Second, another objective of the project was to identify the factors most strongly correlated with substance use (both negatively and positively), which could be used to develop more effective programs aimed at discouraging substance use.

There has been a fair amount of research on this problem spanning several decades, including several statistical and machine learning based studies. In general, they have confirmed that a number of personality characteristics and environmental influences are predictive of early initiation to alcohol and other substances and high intensity of use. In particular, several have concluded that individual attitude characteristics and quality of parental relationships are some of the most highly indicative factors for early onset substance use. [1, 2] Over the course of this project, factor analysis was used to examine and verify these claims.

Several challenges are immediately apparent when considering this problem, however. First, the dataset is fairly skewed towards non-use, especially for 'harder' drugs. To overcome this issue, it was decided to restrict predictions to use of alcohol, tobacco, and marijuana, all of which had a significant population of users (~20%). Second, it is obvious that substance use does not stem from a single, or even specific combination of several, factors. This makes use hard to predict, as it is a fairly complex problem, with psychological and environmental components. However, judging by the correlations exhibited by other studies, and the large number of features in the dataset - 264 per participant, as well as their large range of topics, I hoped to achieve some greater degree of accuracy than a naive classifier.

2 Approach

During the project, three methods were tested and compared with the baseline method, naive classification. All of the approaches were basic machine learning algorithms including logistic regression, unweighted linear regression, and kernelized support vector machines. They were chosen for inclusion for various reasons.

Unweighted linear regression is a simple algorithm with a closed form solution. It is a simple maximum likelihood based estimator which can be run very quickly, with a time complexity of $O(mn^2)$ and space complexity of $O(mn)$ where m is the number of training examples and n is the length of the feature vector. It was chosen to contrast with the other two algorithms in both simplicity and low space and time complexity and expected to perform poorly as a result.

Logistic regression is also a maximum likelihood estimator with uses the sigmoid function as a probability distribution. It has a time complexity of $O(mnk)$ and a space complexity of $O(mn)$ where m and n are as defined above and k is the number of iterations. For this project, $k = 10$, which was chosen as both a reasonable choice for convergence and to put the runtime in between linear regression and the support vector machine. It was chosen as an algorithm of mid range simplicity and complexity.

Finally, a kernelized support vector machine (SVM) was evaluated against the other two algorithms. SVM's use hinge loss to maximize the margin between a high dimensional decision boundary and the examples on both sides closest to it. SVM's can expand the feature vector into higher dimensions using kernels, which are a set of functions that compactly represent the dot product of two transformed vectors ie. $K(\vec{x}, \vec{z}) = \phi(\vec{x}) \cdot \phi(\vec{z})$ for some function ϕ. The space and time complexity of the SVM depends on the kernel chosen, but in general, both are higher than the space and time complexity of the other two algorithms. The SVM was chosen as the most complicated and hopefully accurate algorithm.

3 Experiments

For the experiments, the three algorithms were all trained over a randomized sample of 80% of the dataset and tested on the remaining 20%. A binary vector was extracted from the dataset with -1 values indicating no use of the substance and +1 values indicating use for alcohol, tobacco, and marijuana. Initially, the feature vector included the responses to all 264 questions on the survey excluding those for direct use of the substances. Unfortunately, this produced a 100% successful classification rate over the dataset, due to the inclusion of several, directly correlated features - for example, "use over the past month" - which made the classifier's job far too simple. Thus, feature analysis was performed on the survey, and 45 questions which were directly related to substances were removed from the dataset, leaving 219 features to use for classification. These features were organized into the categories of demographics, computer usage, affluence, living arrangements, physical activity, nutrition, body image, hygiene, physical and mental health, interpersonal relationships, school activity, bullying, and parents' occupations.

After the feature vector had been trimmed, linear and logistic regression were run over the dataset, producing the following results (to see the code, please read the attached files):

Baseline Results:	true pos.	false pos.	true neg.	false neg.	accuracy
Marijuana	0	0	1378	339	80.484%
Alcohol	0	0	1410	326	81.106%
Tobacco	0	0	1369	409	76.997%

Linear Regression:	true pos.	false pos.	true neg.	false neg.	accuracy
Marijuana	335	1090	308	4	37.018%
Alcohol	319	1084	326	7	37.154%
Tobacco	402	1094	275	7	38.076%

Logistic Regression:	true pos.	false pos.	true neg.	false neg.	accuracy
Marijuana	168	74	1324	171	85.895%
Alcohol	170	78	1332	156	86.521%
Tobacco	191	101	1268	218	82.058%

As you can see, the results of linear regression were fairly dismal, while logistic regression managed to outscore the baseline in all cases by around 5.5%. Its runtime, however, was fairly longer, as expected.

To further analyze the results of linear regression, it is clear that the classifier was skewed towards a positive prediction. Had I wanted to further optimize the performance of this algorithm, I could have transformed the dataset into a more symmetric form and rerun the algorithm. However, the linear regression algorithm was not really my focus, more a yardstick for the performance of the other two, so I moved on.

Logistic regression produced a well trained classifier. In the end, it turned out to have nearly the accuracy of the SVM with tuned hyperparameters. Moreover, the algorithm had a significantly shorter runtime than the SVM with the chosen number of iterations. Had I wanted to optimize its performance further, I could have increased the number of iterations and analyzed the training dataset to remove outliers. The theta vector was also analyzed for the features with the highest weights, corresponding to the features with the highest correlation for or against substance abuse. This list broadly confirmed the findings of other, similar studies in terms of most important risk factors. The full analysis is in the conclusion.

Finally, the SVM was run against the training data. The LIBSVM library version 3.14 was used to both train the model and test it over the test set. The data was output from Matlab into a LIBSVM compatible format and the algorithm was run in the shell. Initially, no hyperparameters or optimization techniques were used to provide a baseline for further optimization.

```
./svm-train trainingData model
./svm-predict testingData model output
```

	Marijuana accuracy	Alcohol accuracy	Tobacco accuracy
Initial SVM Results:	80.484%	81.106%	76.997%

As you may notice, these results are identical to those of the naive classifier, indicating that the model always predicted -1. Obviously, this was an overfitting of the skewed data, so it was decided to scale the feature vector in order to reduce variance. Then, the test was run again, with the following results:

```
./svm-scale -s scale_params trainingData > scaleTrainingData
./svm-scale -r scale_params testingData > scaleTestingData
./svm-train scaleTrainingData scaleModel
./svm-predict scaleTestingData scaleModel scaleOutput
```

	Marijuana accuracy	Alcohol accuracy	Tobacco accuracy
Scaled Data SVM Results:	83.592%	84.447%	80.8221%

At this point, several options for optimization manifested themselves. First, there were several kernels to train and test the dataset on. Second, from the output file it was clear that the model was still skewed towards a -1 prediction, so another idea for optimization was to weight the +1 examples more heavily than the -1 using a weight hyperparameter. Finally, once kernels had been selected, there were additional hyperparameters to optimize for each.

To examine the effect of different kernels, the SVM was run again with linear, polynomial (of degree 3), sigmoid, and radial basis function kernels. In the interest of time, it was decided to select the two highest performing kernels for further optimization, which turned out to be the linear and radial basis function kernels.

Next, the hyperparameters of each kernel were tuned to optimize accuracy. The linear function takes a cost parameter and the radial basis function kernel takes both a cost parameter and a

gamma value. Using an exponentially growing test for the linear cost and a an exponential grid search to optimize radial basis function cost and gamma, the optimal values of each were ascertained. Unfortunately, there is not room in my writeup to include all of these figures, but ultimately, the optimal cost for the linear kernel stood at $c = 2^0$ while the optimal parameters for the radial basis function were determined to be $c = 2^5, \gamma = 1/(2^7)$. These parameters turned out to be optimal for all three datasets, strangely, but likely because they were fairly similar.

Finally, in recognition of the remaining skew towards a positive prediction, the weight values were optimized for both kernels. Binary search over the -4 to +4 range weighing the +1 to the -1 values converged on the optimal weight value for each dataset, which did differ between the three substances. Ultimately, the following optimal accuracies were produced, including each algorithm:

	Marijuana accuracy	Alcohol accuracy	Tobacco accuracy
Linear SVM	86.816%	86.233%	82.621%
Optimal Weight	1.5	0.75	1.25
Radial Basis Function SVM	85.607%	85.369%	81.159%
Optimal Weight	0.5	0.75	0.75
Logistic Regression	85.895%	86.521%	82.058%
Linear Regression	37.018%	37.154%	38.076%
Baseline	80.484%	81.106%	76.997%

As you can see, the tuned linear kernel turned out to be the most effective algorithm overall, although its performance was narrowly exceeded by the logistic regression algorithm when predicting alcohol use. Interestingly, the optimal +1 weight parameter for the alcohol use dataset for the linear kernel turned out to be less than 1, which was unexpected, due to the preponderance of -1 examples.

4 Conclusion

Overall, the results of the project were somewhat disappointing, although somewhat informative. The best algorithm did clearly exceed the baseline test, but not by an enormous margin. It is likely that the complexity of the social problem examined (substance abuse) lends itself to a difficult prediction, since it would seem more than possible for two teens with the same lifestyle habits to make individual choices whether or not to abuse a substance. However, the relative success of the algorithm over the baseline illustrates that lifestyle features are correlated with substance abuse in teens in some way, which supports the existing literature. In particular, to return to the analysis of the results of the logistic regression, the highest common risk factors of the three substances consisted of more friends using substances, older age, and separation of parents, while the factors most correlated to lack of abuse consisted of asian heritage, white heritage, living with one's grandmother, and interestingly, living in foster care. These results confirm those of previous studies with the exception of the influence of friends using substances, which previous studies have claimed to be secondary to personal traits, but which in my analysis turned out to be by far the most important risk factor for substance abuse.[3]

5 References

References

[1] Marsha E. Bates and Erich W. Labouvie. Adolescent risk factors and the prediction of persistent alcohol and drug use into adulthood. *Alcoholism: Clinical and Experimental Research*, 21(5):944–950, 1997.

[2] Patricia L. Dobkin and Richard E. Tremblay. Individual and peer characteristics in predicting boys' early onset of substance abuse: A seven-year longitudinal study. *Child Development*, 66(4):1198 – 1214, 1995.

[3] J.D. Hawkins, R.F. Catalano, and J.Y. Miller. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: implications for substance abuse prevention. *Psychological bulletin*, 112(1):64, 1992.

[4] Steve Sussman, Clyde W Dent, and Linda Leu. The one-year prospective prediction of substance abuse and dependence among high-risk adolescents. *Journal of Substance Abuse*, 12(4):373 – 386, 2000.