Recidivism of Criminal Offenders

A machine-learning approach to justice

Estienne Granet
egranet@g.harvard.edu
Harvard University

Victor Lei
vlei@g.harvard.edu
Harvard University

Abstract

Recidivism of criminal offenders is a difficult topic in criminal justice. If we view the justice system as serving an important rehabilitatory purpose, then recidivism represents indicators for potential improvement in the correctional system. Using a public data-set on two cohorts of inmates released from North Carolina prisons in 1978 and 1980, we build two predictors of whether an individual will offend again in the future. Our first predictor is one-layer feed-forward neural networks. The second predictor is random forest. Results are presented below.

Introduction

Our goal is to effectively forecast whether an offender will offender again in the future when he or she is released from prison. To achieve this goal, we rely on a survey conducted on prisoners released from North Carolina prisons in 1978 and 1980. The survey contains questions on the background of the offenders, including their involvement in drugs or alcohol, level of schooling, nature of the crime resulting in the sample conviction, number of prior incarcerations and recidivism following release from the sample incarceration. The data collection also contains information on the the marital status, the sex, the age and the race of the offenders. In hard numbers, the dataset contains 9,327 profiles for prisoners released in 1978 and 9,549 profiles for prisoners released in 1980. Each entry has 16 variables and most variables are indicators.

Past work on the topic include an article on recidivism prediction [Palocsay et al.(2000)Palocsay, Wang, and Brookshire]. In the article, Palocsay et al. fit a one-layer neural network model on the data set, thus providing us with a reference for comparison. Historically though, recidivism prediction has been based on explicative models borrowed from survival analysis such as [Chung et al.(2003)Chung, Schmidt, and Witte] or [Schmidt and Witte(1988)].

Neural Network

Neural networks were discovered some time ago, but lately, they have been experiencing a resurgence in usage. There is likely significant complexity in the interactions between the different variables, so simple models like logistic regression are not going to be as capable of modeling them compared to neural networks. Prior work on this data set has considered the use of neural networks and found them to have good predictive power for recidivism compared to traditional modeling techniques like logistic regression models. We seek to harness the improvements in neural networks that have been made over the last few years in an attempt to improve predictive accuracy. We find that we can achieve a small, but notable improvement in accuracy on unseen data, as well as a significant reduction in the number of nodes in the hidden layer.

As this is a plain binary classifications problem, we use a multi-layer perceptron feedforward neural network with a single hidden layer, bias terms, and the logistic function as the activation function (f) for both the hidden and output layers. There is a single output neuron ($\hat{y} \in (0, 1)$) used for classification with a 0.5 threshold. For some input vector x, linear weights W, and bias terms b:

$$\hat{y} = f(W_2 f(W_1 x + b_1) + b_2)$$

$$E = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

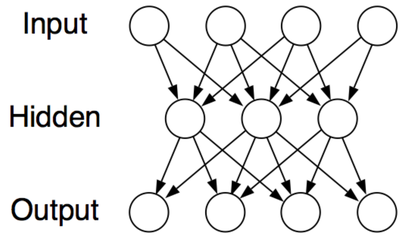

Figure 1: Feedforward Neural Network[Jordan and Bishop(1997)]

Using cross-entropy error (E), we can apply backpropagation and optimize the weights and bias terms by stochastic gradient descent using $\frac{\partial E}{\partial W}$, and $\frac{\partial E}{\partial b}$. In order to counteract overfitting, we use an early-stopping heuristic to stop training the model when the validation set performance does not improve for a certain number of epochs.

After splitting the 1978 dataset and the 1980 dataset into training and validation sets (7:3 ratio) and training the neural network, validation set performance is improved compared to the results from Palocsay et al [Palocsay et al.(2000)Palocsay, Wang, and Brookshire]. The results are also achieved with far fewer nodes in the hidden layer, making it faster to train and likely more generalizable. As the figures below reveal, the neural network starts to quickly overfit as the number of nodes in the hidden layer increases.

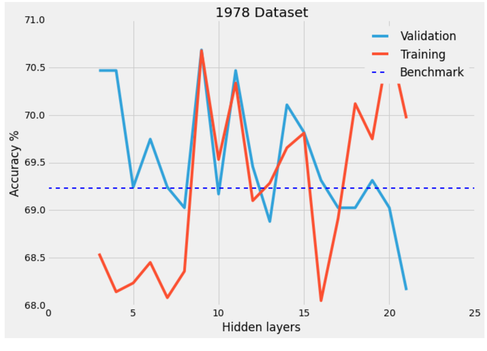

Figure 2: 1978 Dataset

Random Forest

A Random Forest predictor was built over the data in order to get some sense of what features were significantly impacting the odds of recidivism. Two loss functions - a Gini loss and an entropy loss - were tested without any conclusion difference. We simulated over a number of trees ranging from 10 to 100, as higher values were not bringing any noticeable improvement. The number of features tested ranged from 1 to all. The figure below gives an idea of the best features.

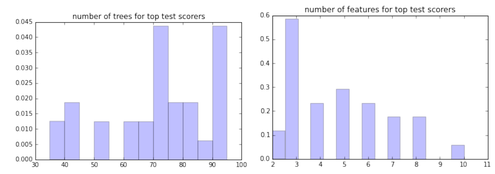

Figure 3: Number of trees and number of features for the top 5% test scores. Simulations were performed for a feature number ranging from 1 to all, a number of trees ranging from 10 to 100, and a Gini loss.

The best result were obtained for a test set reduced to 10% of the initial dataset - in this case, the top test score repeatedly reach 71 % as we shuffle over and over test and training sets. When we use the NN splitting ratio ($7 : 3$), we obtain the following results.

Dataset	Score on test set
1978	67.46 %
1980	67.95 %

Table 2: Best performance for the Random Forest predictor. Test set is identical for NN and Random Forest.

Remarks on the dataset

Score plateau The score on the test set seems to have an upper bound at 72%, independently of the classification method. One explanation of this plateau is the existence of antagonist duplicates in the data set, that is to say two individuals with very similar features but one is a recidivist and the other is not. Such duplicates would structurally lower the score. The existence of duplicates is made easier by the fact that 13 of the 16 feature variable are indicator functions. To address the concern, we looked at the data set. All non-indicator variables were partitioned in groups. For example, the age variable (expressed in months) was replaced with a categorical variable that indicates whether the individual belongs to age group $[0, 19], [20, 24], [25, 29], [30, 34], \ldots$. Depending on how we define these groups, we get a duplicate rate between 1 % and 10 %, which is only a partial explanation for the plateau. Another is simply the lack of data. The 11 indicator variables represent $2^11 = 2048$ combinations. If we add 5 non-indicator categorical variables, we get a a number of combinations that is significantly bigger than $9,327$, the number of data points for prisoners released in 1978 or even the total number of data points if we merge the 1978 and 1980 files.

Explicative variables Random forests classifiers indicate that the number of years spent at school, the age and the duration of the sentence were more important features. This result should be treated cautiously as these three variables happen to be the only 3 variables that are not indicators. Their discriminative power is consequently strengthened.

WHITE	ALCHY	JUNKY	SUPER	MARRIED
0.027672	0.026039	0.030817	0.032681	0.025796
0.028776	0.025371	0.030955	0.032682	0.024949
0.034686	0.024705	0.030558	0.031941	0.025562
PROPTY	**PERSON**	**MALE**	**PRIORS**	**SCHOOL**
0.020870	0.010313	0.012501	0.076603	0.123440
0.019497	0.009766	0.012326	0.077637	0.125512
0.019516	0.008792	0.012404	0.080565	0.117973
TSERVD	**WORKREL**	**AGE**	**FELON**	**RULE**
0.223409	0.031944	0.266270	0.020389	0.071259
0.222114	0.033548	0.260950	0.020042	0.075875
0.227115	0.030774	0.270926	0.017471	0.067011

Figure 4: Importance weights for the top 2 test scores among all classifiers simulated.

Conclusions

We improve upon the existing work done by Palocsay et al. [Palocsay et al.(2000)Palocsay, Wang, and Brookshire] on this data set by achieving a higher validation set accuracy using fewer nodes in the hidden layer. The random forest classifiers indicated that the main factors involved in recidivism are the age, the time served in jail and the level of education.

Forthcoming Research

We are still exploring other ways of optimizing the neural network and more advanced regularization strategies like Dropout. In parallel, we are looking into a hierarchical mixture of models (HME). We hope that a HME will be less prone to local maxima - a frequent problem in our simulations.

References

[Palocsay et al.(2000)Palocsay, Wang, and Brookshire] Susan Palocsay, Ping Wang, and Robert Brookshire. Predicting criminal recidivism using neural networks. *Socio-Economic Planning Sciences*, 34(4):271–284, Decemeber 2000.

[Chung et al.(2003)Chung, Schmidt, and Witte] Ching-Fan Chung, Peter Schmidt, and Ann Witte. Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1):59–98, March 2003.

[Schmidt and Witte(1988)] Peter Schmidt and Ann Witte. *Predicting Recidivism Using Survival Models*. Springer-Verlag, New York, 1988.

[Jordan and Bishop(1997)] Michael Jordan and Chris Bishop. An introduction to graphical models, December 1997.

[Schmidt and Witte(1989)] Peter Schmidt and Ann Dryden Witte. Predicting criminal recidivism using split populationsurvival time models. *Journal of Econometrics*, 40(1):141–159, 1989.

[Murphy(2012)] Kevin Murphy. *Machine Learning, A Probabilist Perspective*. MIT Press, August 2012.