Drug Classification

Riddhi Barbhaiya, Rohit Haritsa, Helen Le

Introduction

Project Description/Summary

The goal of this project is to predict consumption of various drugs(specified below) using personality and demographic information. We treat this as a multi-label classification problem. To solve this classification we explore two methods:

1) One vs All classification: we build five classifiers - one for each drug of interest
2) Neural Network: we build a neural network that predicts all 5 drugs simultaneously

Strategy one does not consider any relations between the drugs which may be informative in classification. On the other hand, due to the dataset being fairly small, the neural network is more likely to overfit. We measure the performance of the two methods by looking at global measures such as hamming loss and subset accuracy. We also evaluate measures per class such as AUC, recall, and specificity.

Dataset

We use the drug consumption (quantified) dataset from the UCI Machine Learning Repository. The dataset includes 12 predictor variables — age, gender, education, country, ethnicity, 7 personality measures. Many of the predictor variables are categorical. The dataset that we use has quantified these variables using ordinal and nominal feature combination and sparse PCA (Fehrman et al, 2017). We focus on predicting usage of five drugs: cocaine, LSD, heroin, benzodiazepine (benzos), and caffeine. Predicting drug consumption can be applicable in predicting abuse and addiction. Demographic information and personality information are easy to obtain and would be a feasible to use to predict abuse.

Results Summary

We are able to classify each drug with over 70% accuracy (table 2). Both methods perform comparably. This indicates that there isn't really an advantage to employing a complicated method(neural network) when the simpler method performs closely.

Methodology

Data Preparation

Data preparation was relatively straightforward. After reading the csv file into Python, we used simple indexing to extract the relevant variables we wanted to analyze. The dataset contains information of the frequency of usage but we convert frequency to users and nonusers. Users are considered people who have used a drug in the last year or more frequently. Non-users are considered people who have used a drug more than a year ago or never. We converted the values such that 0 corresponds to non-users while the 1 corresponds to users of the particular drug.

One-vs-All Logistic Classifier

After extracting the required variables from the dataset, we used One vs All logistic regression model. Since the dependent variable (users or non-users) is binary, logistic regression gave us the most accurate results. Since our final goal is to predict whether an individual will be a user or a non-user of a particular drug, predictive analysis was the best fit.

Neural Network Model

We also fit a neural network model because we wanted to explore a method that would simultaneously predict all the drugs of interest. This would allow the network to pick up on the relationship between the drugs of interest. Cross validation was used to tune hyperparameters and the architecture of the network. The final network's specifications are given in the implementation details section.

Model Evaluation

In a multi-label problem as we consider, there are many ways one's predictions for a sample can be wrong. There are a plethora of measures that can be used to assess performance. There is also interesting theoretical work that shows that not all measures can be maximized together (Dembczynski et al., 2010). In our project, we consider the following measures to evaluate performance:

- ROC/AUC(per drug class): we calculated the area under roc curves to evaluate the classification of each drug separately.
- Hamming loss: We used hamming loss to assess the performance of both the neural network and logistic regression models on all five classes of the drugs. We use hamming loss because it provides a flexible measure of multi-label accuracy and does not enforce that all drug classes have to be predicted correctly.
- Subest accuracy: We also evaluate the performance of the two classification strategies using subset accuracy. This calculates the percentage of individuals with all drugs classified correctly.

Hamming loss and subset accuracy are used as defined by Karthik Nooney[3].

Implementation Detail

One-vs-All Logistic Classifier

For each of the five drugs that we built a classifier for, we split the corresponding dataset so that 75% of the data would train the logistic regression model while 25% of the data was used to test the performance of the model. The same steps were taken to build each model.

We used `LogisticRegression` with `multi_class = "ovr"` to denote that we were building a logistic regression model with one-vs-all multi-classification. For example, benzos would be compared to caffeine, cocaine, LSD, and heroin as a group to predict whether an individual is a user of benzos or not. Then we fit the model with our training data and used it to predict users with our test data. We also looked at the accuracy score for the predictions of all the models.

We also calculated ROC curves to show the performance of the classification model. We calculated the scores of "no skill" predictions vs. with the logistic classifier with our predictions

and our actual test results. We then had our ROC AUC scores to compare. Then we plotted the ROC curves to see the visual representation of our classifier performance based on the true positive rate and the false positive rate of our predictions. We could then see how well our classifier performed.

Neural Network Model

Model selection:

The dataset was split such that 75% of the dataset was used to train and select the neural network and 25% of the dataset was reserved for testing the performance of the model. The size of the hidden layer and the learning rate were chosen by 5 fold cross validation(CV). Sizes from 4 to 20 in steps of 4 were considered for the hidden layer(as recommended by__). For each of the sizes learning rates (0.3,0.1,0.01,0.001) were also considered. For each possible combination- 5 fold CV was used to calculate hamming loss, subset accuracy, and weighted AUC were calculated (averaged over folds). These measures were used because they are easily comparable therefore easier to use for model selection.The combination of parameters that performed the best were then used to train the final model.

Final/selected network specifications

Network architecture: 12 unit input layer, a 4 unit hidden layer, and a 5 unit output layer. ReLu activation was used to transform inputs to the hidden layer and sigmoid activation was used to transform hidden activations to outputs. Activities in the output above 0.5 were considered users while rest were considered non-users.

Training: Stochastic gradient descent was used to minimize binary cross entropy loss of the network. The learning rate was 0.3. Batch size was set to 32 and 200 epochs were used.

Testing: The network was trained using a training set as defined above. Performance was evaluated on the test set.

Results/Interpretation

One-vs-All Logistic Classifier

All of our AUC scores for each of the classifiers (benzos, caffeine, cocaine, LSD, and heroin) were above 0.650, which means that all of the models are better at predicting true values than “no-skill” predictions. (All of the scores can be found in Table 1.) The models are then good at distinguishing between positive and negative classes. All of our ROC curves, shown in Figures 2-6, demonstrate the probabilities of our predictions and indicate that our classifiers have higher true positive rates than false positive rates. Thus, we can say our one-vs all logistic classifiers for benzos, caffeine, cocaine, LSD, and heroin users vs. non-users perform well. The hamming loss tells us that only 15.3% of all the labels in all the drugs are misclassified. The subset accuracy tells us that only 51.3% of the individuals have all their drugs classified correctly.

Neural Network Model

Most of the AUC scores for the neural network model are comparable to the logistic regression classifiers. This indicates that for most drugs, the network is better at predicting the

true values than random guessing. Interestingly, the caffeine classification performance is the worst. This may be because a large proportion of individuals are considered caffeine users so the network has difficulty distinguishing between users and non-users. The hamming loss tells us that only 15.6% of all the labels in all the drugs are misclassified. The subset accuracy tells us that only 52% of the individuals have all their drugs classified correctly.

Comparison of the two methods

Over all, the logistic regression models applied to each drug and the neural network model perform similarly. We had hypothesized that the NN model may do better because it is trained on all drugs simultaneously. Since the neural network model does not perform better than individual regression models, there either isn't much useful correlation between drugs or the network isn't able doesn't learn to use it.

There are also some interesting cases that occur in both the models. In predicting caffeine both models do not predict any individual to be a non-user. This can be explained by the ubiquity of caffeine usage. Additionally, in predicting Heroin usage, the models do not predict any individual to be a user. This may be because heroin consumption in this data set is quite low. The proportion of usage for each drug is shown in figure 1 and table 2 shows the measures that indicate the two cases discussed above.

Overall

In doing this project, we learned how to deal with a multi-label problem. One strategy we used was to break it down into binary classification problems that we knew how to solve. We understand that using this method does not take into account any relationships between the drugs we are trying to predict. To address this concern we also used a neural network model that predicted all the drugs simultaneously. The hope was that this would allow the network to pick up on relationships between the drugs. Our results indicate that this was not the case. To exploit the relationships between drugs, in the future one can explore chaining classifiers and structured learning methods.

Additionally, we realized that evaluating a multi-label problem is not as straightforward. Especially how to average over classes is a challenge. In the end, we settled on three performance measures specified above. We learned how to use AUC-ROC to measure the performance of the classifiers. This was nice to learn because it seems to be broadly applicable. Lastly, we also explored neural networks models and learned how to tune hyperparameters using cross validation.

In all, the performance of our classifiers is comparable to the investigation done by Fehrman et al. By conducting this investigation we learned to deal with a multi-label classification problem.

Supplementary Materials

Figure 1

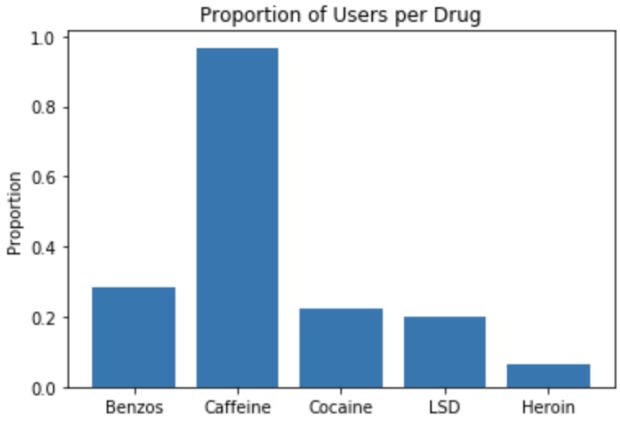

Figure 2

Figure 3

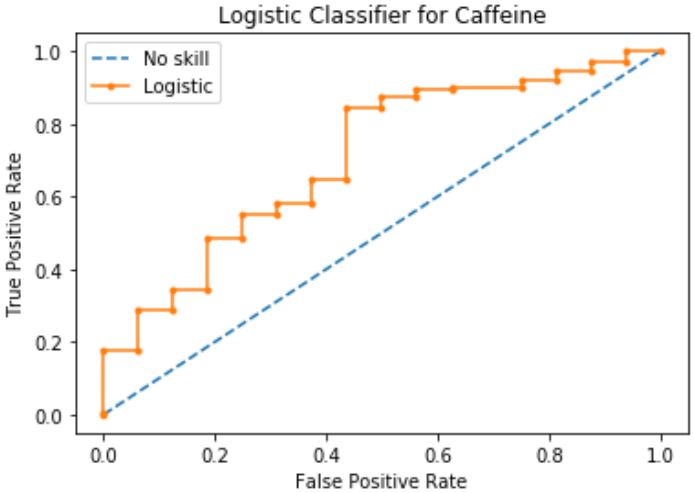

Figure 4

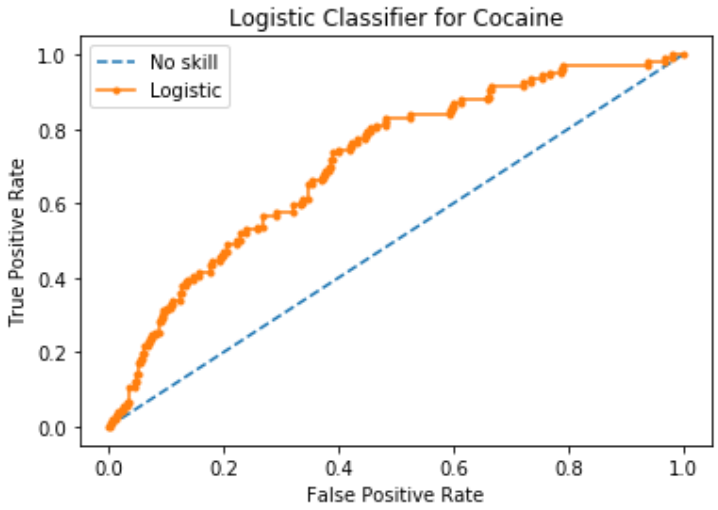

Figure 5

Figure 6

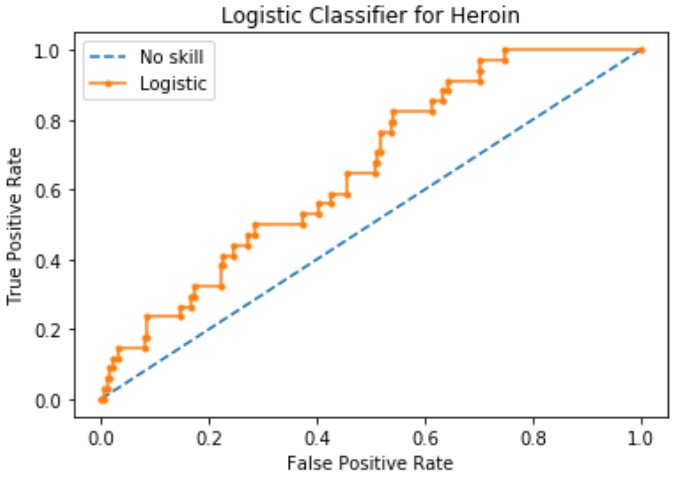

Table 1: AUC of logistic regression models and neural network(NN) model

Drug	LogReg	NN
Benzos	0.766	0.757
Caffeine	0.708	0.593
Cocaine	0.712	0.759
LSD	0.700	0.896
Heroin	0.659	0.794

Table 3: Performance measures for the two models

Errors	Log Reg	NN
Hamming	0.158	0.156
Subset Accuracy	0.513	.520

Table 2: Accuracy, recall, and specificity of each drug in Logistic Regression and NN models

Drug	Accuracy	Recall	Specificity	Accuracy	Recall	Specificity
Benzos	0.728	0.394	0.873	0.732	0.422	0.857
Caffeine	0.966	1	0	0.970	1	0
Cocaine	0.782	0.179	0.956	0.764	0.207	0.936
LSD	0.805	0.495	0.887	0.821	0.416	0.932
Heroin	0.928	0	1	0.930	0	1

References

Dataset: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#

[1] Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN (2017) The five factor model of personality and evaluation of drug consumption risk. In: Palumbo F, Montanari A, Vichi M. (eds) Data science. Studies in classification, data analysis, and knowledge organization. Springer, Cham

[2]Dembczynski, Krzysztof, Waegeman, Willem, Cheng, Weiwei, and Hu ̈llermeier, Eyke. Regret analysis for per- formance metrics in multi-label classification: the case of hamming and subset zero-one loss. In ECML/PKDD, pp. 280–295. 2010.

[3] Nooney, Kartik. “Deep dive into multi-label classification..! (With detailed Case Study)” Towards Data Science, 7 Jun. 2018. Web. 19 March 2009.