Prediction of Autism

Fateme Kamiab
7 January, 2020

1. Introduction

1.1 Background

Autism spectrum disorder is one of the most commonly reported mental and behavioral disorder in children. it is a condition or impairment of central nervous system function, which results in insufficiency of the information reaching the brain. In the fourth Diagnostic and Statistical Manual of Mental Disorders, Autism spectrum disorder has been introduced as a subset of developmental disorders in which there is a disorder in social skills, behavior, and language development. Among all developmental disorders, Autism has the highest prevalence and constitutes more than two thirds of these disorders.

1.2 Problem

Due to the complexity, variety and the closeness of the symptoms of Autism and other developmental disorders, the diagnosis of Autism is not easily possible. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of Autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis

2. Data acquisition and cleaning

2.1 Data Source

The rapid growth in the number of ASD cases worldwide necessitates datasets related to behavior traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited Autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to Autism screening of toddlers that contained influential features to be utilized for further analysis especially in determining Autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioral features (Q-Chat-10) plus other individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behavior science.

2.2 Data preparing

Data Type: Predictive and Descriptive: Nominal / categorical, binary and continuous
Task: Classification
Attribute Type: Categorical, continuous and binary
Area: Medical, health and social science
Format Type: Non-Matrix

Does your data set contain missing values? No
Number of Instances (records in your data set): 1054
Number of Attributes (fields within each record): 18 including the class variable
Attribute Information: For Further information about the attributes/feature see below table.

A1-A10: Items within Q-Chat-10 in which questions possible answers: "Always, Usually, Sometimes, Rarely & Never" items' values are mapped to "1" or "0" in the dataset. For questions 1-9 (A1-A9) in Q-chat-10, if the response was Sometimes / Rarely / Never "1" is assigned to the question (A1-A9). However, for question 10 (A10), if the response was Always / Usually / Sometimes then "1" is assigned to that question. If the user obtained More than 3 Add points together for all ten questions. If your child scores more than 3 (Q-chat-10- score) then there is a potential ASD traits otherwise no ASD traits are observed.

Table 1: Details of variables mapping to the Q-Chat-10 screening methods

Variable in Dataset	**Corresponding Q-chat-10-Toddler Features**
A1	Does your child look at you when you call his/her name?
A2	How easy is it for you to get eye contact with your child?
A3	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)
A4	Does your child point to share interest with you? (e.g. poin9ng at an interes9ng sight)
A5	Does your child pretend? (e.g. care for dolls, talk on a toy phone)
A6	Does your child follow where you're looking?
A7	If you or someone else in the family is visibly upset, does your child show signs of wan9ng to comfort them? (e.g. stroking hair, hugging them)
A8	Would you describe your child's first words as:
A9	Does your child use simple gestures? (e.g. wave goodbye)
A10	Does your child stare at nothing with no apparent purpose?

Also, we discard the Score variable as it has been used to assign the class label so if you keep the score variable the models derived might be overfitted.

Table 2: Features collected and their descriptions

Feature	Type	Description
A1: Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A2: Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A3: Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A4: Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A5: Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A6: A6: Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A7: Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A8: Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A9: Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A:10 Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Age	Number	Toddlers (months)
Score by Q-chat-10	Number	1-10 (Less that or equal 3 no ASD traits; > 3 ASD traits
Sex	Character	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with ASD history	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Why_are_you_taken_the_screening	String	Use input textbox
Class variable	String	ASD traits or No ASD traits (automatically assigned by the ASDTests app). (Yes / No)

This research used binary, categorical and continuous measures, mainly from samples, to craft an extensive list of explanatory variables comprised of various elements measured within a person's behavior. The goal was to find a "best-fitting" model using logistic regression described below. The ability to predict the probability (risk) of a child having ASD through some traits could greatly contribute to the difficult diagnosis process currently in place.

3. logistic regression

While Linear Regression is suited for estimating continuous values (e.g. estimating house price), it is not the best tool for predicting the class of an observed data point. In order to estimate the class of a data point, we need some sort of guidance on what would be the **most probable class** for that data point. For this, we use **LogisticRegression**.

As you know, **Linear regression** finds a function that relates a continuous dependent variable, **y**, to some predictors (independent variables x_1, x_2, etc.). For example, Simple linear regression assumes a function of the form:

$$y=\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots$$

and finds the values of parameters θ_0, θ_1, θ_2 etc, where the term θ_0 is the "intercept". It can be generally shown as:

$$h_\theta(x)=\theta^T X$$

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, **y**, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

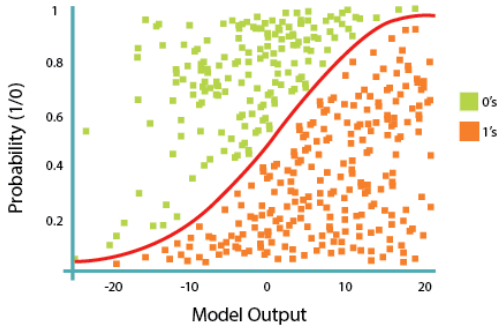
Logistic regression fits a special s-shaped curve by taking the linear regression and transforming the numeric estimate into a probability with the following function, which is called sigmoid function σ:

$$h_\theta(x) = \sigma(\theta^T X) = \frac{e^{(\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots)}}{1+e^{(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots)}}$$

$$ProbabilityOfaClass_1 = P(Y=1|X) = \sigma(\theta^T X) = \frac{e^{\theta^T X}}{1+e^{\theta^T X}}$$

In this equation, $\theta^T X$ is the regression result (the sum of the variables weighted by the coefficients), exp is the exponential function and $\sigma(\theta^T X)$ is the sigmoid or logistic function, also called logistic curve. It is a common "S" shape (sigmoid curve).

So, briefly, Logistic Regression passes the input through the logistic/sigmoid but then treats the result as a probability:



The objective of **Logistic Regression** algorithm, is to find the best parameters θ, for $h\theta(x) = \sigma(\theta^T X)$, in such a way that the model best predicts the class of each case.

4. Modeling (Logistic Regression with Scikit-learn)

Let's build our model using **Logistic Regression** from Scikit-learn package. This function implements logistic regression and can use different numerical optimizers to find parameters, including 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga' solvers.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
LR = LogisticRegression(C=0.01, solver='liblinear').fit(X_train,y_train)
LR
```

```
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='liblinear', tol=0.0001, verbose=0,
                   warm_start=False)
```

The version of Logistic Regression in Scikit-learn, support regularization. Regularization is a technique used to solve the overfitting problem in machine learning models. **C** parameter indicates **inverse of regularization strength** which must be a positive float. Smaller values specify stronger regularization.

5. Performances of different Evaluation

5.1 Jaccard index

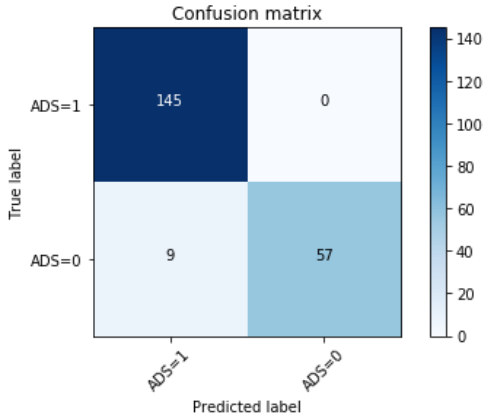
Lets try jaccard index for accuracy evaluation. we can define jaccard as the size of the intersection divided by the size of the union of two label sets. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0.

```
from sklearn.metrics import jaccard_score
jaccard_score(y_test, yhat)
```

```
0.9415584415584416
```

5.2 Confusion matrix

Another way of looking at accuracy of classifier is to look at **confusion matrix**.



Look at first row. The first row is for Samples whose actual ADS value in test set is 1. As you can calculate, out of 211 samples , the ADS value of 145 of them is 1. And out of these 145, the classifier correctly predicted 145 of them as 1, and 0 of them as 0.

It means, for 145 samples, the actual ADS value were 1 in test set, and classifier also correctly predicted those as 1.

What about the samples with ADS value 0? Let's look at the second row. It looks like there were 57 customers whom their ADS value were 0.

The classifier correctly predicted 57 of them as 0.and 9 of them wrongly as 1 So, it has done a good job in predicting the samples with ADS value 1. A good thing about confusion matrix is that shows the model's ability to correctly predict or separate the classes. In specific case of binary classifier, such as this example, we can interpret these numbers as the count of true positives, false positives, true negatives, and false negatives.

5.3 Classification's Report

```
print (classification_report(y_test, yhat))
```

```
              precision    recall  f1-score   support

           0       1.00      0.86      0.93        66
           1       0.94      1.00      0.97       145

    accuracy                           0.96       211
   macro avg       0.97      0.93      0.95       211
weighted avg       0.96      0.96      0.96       211
```

Based on the count of each section, we can calculate precision and recall of each label:

- **Precision** is a measure of the accuracy provided that a class label has been predicted. It is defined by: precision = TP / (TP + FP)
- **Recall** is true positive rate. It is defined as: Recall = TP / (TP + FN)

So, we can calculate precision and recall of each class.

F1 score: Now we are in the position to calculate the F1 scores for each label based on the precision and recall of that label.

The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It is a good way to show that a classifer has a good value for both recall and precision.

And finally, we can tell the average accuracy for this classifier is the average of the F1-score for both labels, which is 1.0 in our case.

5.4 log loss

Now, let's try **log loss** for evaluation. In logistic regression, the output can be the probability of samples with ADS is yes (or equals to 1). This probability is a value between 0 and 1. Log loss (Logarithmic loss) measures the performance of a classifier where the predicted output is a probability value between 0 and 1.

```
from sklearn.metrics import log_loss
log_loss(y_test, yhat_prob)
```

```
0.25636419958456347
```

6. Conclusion

Overall, this exploratory study has shown that there is potential for an adequate model based on predictors used to diagnose ASD in the future. This study was especially limited due to sample size, and even a large sample would still require a screening of the extensive list of predictor variables to identify potentially "important "predictors. Study of the various elements in the sample's behavior with a clinical professional is important in terms of finding an adequate starting point with which to begin model selection. The ability to obtain a larger sample would allow for the consideration of more predictors into the multivariate model.

7. Future work

The information gained in this study provides a basis from which to move forward with further research. Though the model obtained worked well for the current sample, there is a possibility that the variation observed within this sample is not a good representation of the population as a whole. A first step toward future work would be to validate this model repeatedly for other sample data to see how well it performs. Second, and possibly most important, would be to increase the sample size if possible so the model is not as sensitive to influential data and over-fitting. Outliers and influential data had much more of an impact in this study in regards to decision making to preserve sample size. Lastly, an in depth discussion with a clinical professional about the clinical importance of the variables both in the multivariate model and even those excluded from the model could have a significant impact in the model building process. The ability to provide a model for the prediction of ASD would have a huge impact on the current diagnosis procedures and this study provided a first look into how our behavior may provide the clues necessary to solve the problem.