Predicting Depression Using RNA

Ziv Lautman lautman@stanford.edu
Harvey Wang hawang97@stanford.edu
Shannon Xiao sxiao1@stanford.edu

Summary

Mental health issues, such as depression, are a serious concern in our society today and a key factor in tackling this issue lies in its diagnosis. We wanted to investigate the use of machine learning models to analyze RNA sequencing data to help with depression diagnosis. We experimented with several algorithms such as PCA, logistic regression, Support Vector Machine (SVM), neural networks, and anomaly detection to fit our data. Our goal was to determine a classification model for our RNA samples that could predict depression. We found the anomaly detection algorithm worked the best on our test set.

Dataset

As part of one of our group member's research at Dr. Michael Snyder's lab, we obtained a proprietary dataset of 124 samples of RNA sequences, 26 of which were classified as depressed. Each row was an individual sample and each column contained the measurements of the expression of a specific gene. The last column represented the BDI scores (depression scores), which were the ground truth. Since the scores were on a scale from 0-60, we adjusted the output values to fit a binary scale of 0/1; samples with a BDI at/above 14 were considered depressed and labeled 1.0, while the rest were labeled 0.0. We also normalized the input data with feature scaling and mean normalization.

Models

PCA
Using Python's scikit-learn library, we applied PCA to see if we could spatially separate the depressed and non-depressed data.

Logistic Regression
With logistic regression, we fit a logistic function as the hypothesis function $h_\theta(x)$ and learned weights that minimized the cost function:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

SVM
Using Python's scikit-learn library, an SVM with a linear kernel was used to see if it performed better than logistic regression.

Neural Network
A neural network was also trained, with the number of hidden layers and nodes determined by experimentation.

Anomaly Detection
The dataset was modeled as a Gaussian distribution:

$$p(x) = \prod_{j=1}^{n} p\left(x_j; \mu_j, \sigma_j\right) = \frac{1}{\sqrt{2\pi}\sigma_j} exp\left(-\frac{\left(x_j-\mu_j\right)^2}{2\sigma_j^{\,2}}\right)$$

A probability p lower than a threshold value was labeled as anomalous.

Features

The features we used were the gene expressions from the raw data. We compared the genes that Zhao et al. considered to be highly correlated with depression, and identified 98 of their 137 genes within our dataset. We pruned our original dataset to match the genes from Zhao et al. and resulted with an input matrix with 98 feature columns (genes) and one output column for the depression score.

Discussion

Our results showed that our anomaly detection algorithm performed the best on the test set with the highest F1 score of 0.78. While this is a relatively good F1 score, the general results of our experiments were not great, since ideally, we wanted scores > 0.85. Additionally, when testing our logistic model, we were initially producing mostly invalid F1 scores (due to zero division) or scores of 0 and it seems that using 98 features overfit the training data. Thus, we decided to take a subset of 20 genes from the set of features and were able to get better results.

As our dataset size was quite small and skewed, we believe that obtaining a larger dataset in the future could yield better accuracy results. By collecting more data, we could also train an algorithm that will perform multi-class classification based on the BDI scores. With more time, we could also refine our algorithms further, e.g. spending more time tuning the parameters.

Results

The training set, cross-validation set, and test set were divided up as 60%, 20%, and 20% of the complete dataset, respectively. They were randomly divided but retained a proportional split of the two classes, with the training set having 60 non-depressed and 14 depressed samples and the validation and test sets each having 19 non-depressed and 6 depressed.

Model	Hyperparameters	Test F1
Logistic Regression	$\lambda = 0.32$	0.44
SVM	$C = 0.005$	0.43
Neural Network	3 layers (98, 48, 1)	0.53
Anomaly Detection	$\varepsilon = 6.5\times10^{-49}$	0.78

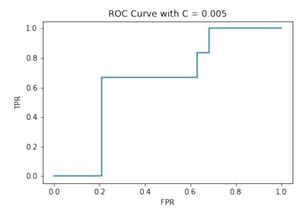

Figure 1: ROC curve for SVM

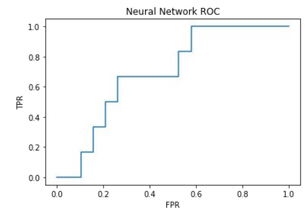

Figure 2: ROC curve for the neural network

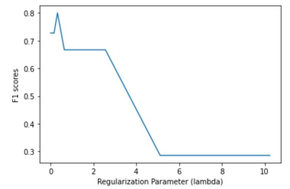

Figure 3: Plot of the F1 scores of the validation set (using the subset of features) against lambda values

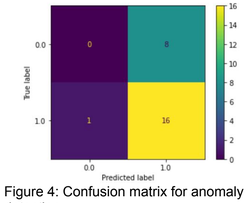

Figure 4: Confusion matrix for anomaly detection

References

[1] National Alliance on Mental Illness. 'Mental Health By the Numbers'. Feb 2022. [ONLINE]: https://www.nami.org/mhstats
[2] Shu Zhao, Zhiwei Bao, Xinyi Zhao, Mengxiang Xu, Ming D. Li, and Zhongli Yang. 'Identification of diagnostic markers for major depressive disorder using machine learning methods'. Frontiers in Neuroscience, 15, 2021.
[3] Source code (requires Stanford login): https://code.stanford.edu/lautman/cs129