Predict Optimized Treatment for Depression CS229 Project Report

Minakshi Mukherjee: adaboost@stanford.edu
Suvasis Mukherjee : suvasism@stanford.edu

I. *Abstract*

For the past 60 years, the anxiety and depression medications are prescribed to patients based on The Hamilton Depression Rating Scale (HDRS)[6] and Social and Functioning Assessment Scale(SOFAS)[2]. The HDRS[6] does not take into account the neuro biomarkers as it is very expensive to do FMRI on all patients. Goal of this project is to identify whether HDRS score and SOFAS score are representative of the three antidepressants: Sertraline, Venlafaxine, Escitalopram prescribed based on FMRI data of 5 brain attributes based on the small dataset of 128 patients collected from Williams Pan-Lab, Precision Psychiatry and Translational Neuroscience, Stanford Medicine iSPOT-D project. There is a need for markers that are predictive of remission and guide classification and treatment choices in the development of a brain-based taxonomy for major depressive disorder (MDD) that affect millions of Americans.

II. *Introduction*

Healthcare professional prescribes antidepressant medications to patients based on two scores: HDRS[6] and SOFAS[2] that are subjective in nature as it is done solely by interviewing the patients. It is very expensive to collect FMRI brain scan data for individual patients to derive a scientific data driven assesment of the patient's medication. We got motivated by this small dataset of 128 patients which contains antidepressants administered to the patients based on HDRS and SOFAS; it also contains FMRI data for 5 important regions from brain. We analyzed this dataset to understand the association between the antidepressants and 5 brain attributes. We tested several models, the input and output are detailed here:

Logistic Regression:
input Dependent Variable: HDRS_response(0 or 1)
input Independent Variable: age,gender,education,3 antidepressants(1 or 0 based on which one patient is taking), 5 brain scan attributes. input comes from training samples.
output: $y.pred = if\ else(prob.pred > 0.5, 1, 0)$
Similar analysis has been carried out with Dependent variable as SOFAS_response and the same set of independent variables.

Linear Regression:
input Dependent Variable: HDRS_baseline(a number less than 100)
input Independent Variable: age,gender,education,3 antidepressants(1 or 0 based on which one patient is taking), 5 brain scan attributes. input comes from training samples.
output: $y.pred =$ HDRS_score for a new x_*
Similar analysis has been carried out with Dependent variable as SOFAS_score and the same set of independent variables and for Ridge regression,Lasso and Elastic Net, which is same as Bayesian Linear Regression with Laplace Prior.

SVM:
input : $y_i f(x_i)$ where y_i denotes HDRS_response(0 or 1) or SOFAS_response(0 or 1) and x_i denotes age,gender,education,3 antidepressants(1 or 0 based on which one patient is taking), 5 brain scan attributes. input comes from training samples.
output: set of weights w (or w_i), one for each feature, whose linear combination predicts the value of y.

Factor Analysis:
input : All 13 feature variables associated to the patient
output: Factor loadings representing the importance of each feature.

Gaussian Mixture Model:
input : All 13 feature variables associated to the patient and the no of components(=4, in our case)
output: Labels of Gaussian Mixture(0,1,2,3 in our

case)

III. *Related Work*

The dataset is small, so we looked into relevant papers that discusses prediction techniques for small dataset. The paper "Regression Shrinkage and Selection via the Lasso" by Robert Tibshirani[1] demonstrated how Lasso enjoys some of the favourable properties of both subset selection and ridge regression and produces interpretable models like subset selection that exhibits the stability of ridge regression. We used this approach to pinpoint the exact brain scan attribute associated to a particular antidepressant. We enhanced the approach by considering Bayesian Linear regression with Laplace prior. "Finite mixture models" by G.J. McLachlan et al.[3] discusses innovative ideas for Bayesian Approach to Mixture Analysis, Mixtures with Non normal Components. In future, we like to incorporate some of these ideas, but in this project we limited ourselves to Gaussian Mixture Model. The paper "Predicting Inpatient Discharge Prioritization With ElectronicHealth Records" by Anand Avanti et al.[5] discusses an extensive use of ensemble classifiers, ROC etc, we implemented similar approach in our project. The paper "Countdown Regression: Sharp and Calibrated Survival Predictions" by Anand Avanti et al.[7] provides ideas about scoring rule as a measure of the quality of a probabilistic forecast. In future, we would like to come up with a scoring mechanism to predict antidepressant based on FMRI data.

IV. *Data Set and Features*

The small dataset has 128 patient IDs and 13 attributes:
.One of three antidepressants taken by them
.Age,gender,years of education
.HDRS score
.SOFAS(Social and Functioning Assessment Scale) score
5 attributes from Amygdala cluster,Insula clusters and Nac clusters.

Since this is a very small dataset with just 128 patient information, we need to employ a few algorithms that fit small data set better. As part of exploratory data analysis, we will build a few supervised and unsupervised models. Unsupervised model will help to understand the similarity among the brain attributes obtained from MRI images and we can use this prior information to build supervised model in order to associate connections between antidepressants and 5 brain attributes.

Our medical data is scarce, so we need a method to make sure the model trained on this dataset will predict with similar accuracy on new patients.

We split the dataset as follows:
We kept 20 percent dataset aside for test and 80 percent for training and validation set. We use K-fold cross validation with K=10.

1. Randomly split dataset S into k disjoint subsets of $\frac{m}{k}$ data in each:$\{S_1, S_2, .., S_k\}$
For each j = 1..k
2. Train model M_i on $S_1 \cup ... \cup S_{j-1} \cup S_{j+1} ... \cup S_k$ and get hypothesis h_{ij}
3. Test hypothesis h_{ij} on S_j and get $\hat{\epsilon}_{s_j}(h_{ij})$
4. Error E_i of Model $M_i = \frac{1}{j}\sum_j \hat{\epsilon}_{s_j}(h_{ij})$
5. Pick model M_i with lowest error E_i
6. Retrain M_i on entire dataset S.
7. Result the hypothesis as the final answer.

Models considered for the project

Logistic Regression
Linear Regression
Bayesian Linear Regression with Laplace Prior
Factor Analysis
Gaussian Mixture Model
SVM

V. *Methods and Algorithm*

Mixture Model

In order to understand the association between HDRS/SOFAS score and the structure of each of the brain scan data, we deep dive further using Mixture Models.

Assumption

A distribution f is a mixture of K component distributions $f_1, f_2, ...f_K$ if
$f = \sum_{i=1}^{K} \lambda_k f_k$.
λ_k are the mixing weights, $\lambda_k > 0$, $\sum \lambda_k = 1$ Here we assume, $f_1, f_2, ...f_K$ follow Gaussian. In the

above, f $\in$ a complete stochastic model, first we pick a distribution, with probabilities given by the mixing weights, and then generate one observation according to that distribution.
Symbolically,

$Z \sim Mult(\lambda_1, \lambda_2, ..., \lambda_K)$
$X|Z \sim f_Z$

We ran different Gaussian Mixture models using HDRS/SOFAS baseline score and brain data.

Factor Analysis

Factor Analysis works on small dataset where it helps to captures the correlations in the data.
Assumption:
dataset $x^{(i)}$ is generated by sampling a k dimension multivariate Gaussian $z^{(i)}$, a latent random variable; $k < 13$, where is 13 is the no of features in our dataset. We like to model the dataset with a joint distribution $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$
$z \sim \mathcal{N}(0, I)$
$\epsilon \sim \mathcal{N}(0, \Psi)$
$\epsilon \; and \; z \; are \; independent.$
$x = \mu + \Lambda z + \epsilon$
$x^{(i)} \; has \; the \; covariance \; \Psi \; noise$
$\mu + \Lambda z \; is \; the \; K - dimensional \; affine \; subspace \; of R^n.$
1.Given the guesses for z that the E-step finds, M step estimates the unknown linearity Λ relating the $x's$ and $z's$.
2.In the final M-step update for Λ, it captures the covariance $\Sigma_{x^{(i)}|z^{(i)}}$ for the posterior distribution $p(x^{(i)}|z^{(i)})$.
3.We declare the convergence when the increase in likelihood $l(\Lambda)$ in successive iterations is smaller than the tolerance parameter.
4.We choose the maximum of $l(\Lambda)$, out of all obtained by k-fold CV.

Bayesian Regression with Laplace Prior

We choose a Laplace prior for the parameter θ. The idea behind choosing Laplace prior is that Laplace distribution is symmetric around zero and it is more strongly peaked as λ grows.
Assumption:
1.σ^2 is known
2.All θs are independent with Laplace density.
3.With this prior, the MAP estimator is the same as the lasso solution, this sparse solution is useful because we have five feature variables for Brain structure, and we would like to establish the functional connectivity between antidepressants and brain structure, so we would like to have some of the θ's zero.

$$Laplace \; Prior : p(\theta) = \frac{\lambda}{2 * \sigma} exp(-\frac{\lambda|\theta|}{\sigma})$$
$$Dataset : \; S = \{x^{(i)}, y^{(i)}\}_{i=1}^m$$
$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$
$$epsilon^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$$

4.We search for a choice of θ that minimizes the objective function

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2$$

5.The output of Bayesian linear regression on a new test point x_* is the posterior predictive distribution $p(y_*|x_*, S) = \int_\theta p(y_*|x_*, \theta)p(\theta|S)d\theta$

$$Parameter \quad Posterior \quad p(\theta|S) = \frac{p(\theta)\prod_i p(y^{(i)}|x^{(i)},\theta)}{\int_{\theta'} p(\theta)\prod_i p(y^{(i)}|x^{(i)},\theta)d\theta'}$$

VI. *Results and Findings*

We built several models using several variations of feature variables from our small dataset. Since dataset is small, it is to our advantage that we can iterate several algorithms as well as permutations of several functions of the feature variables to pinpoint which one improve accuracy of the prediction.

Average test set misclassification error based on validation set is chosen as a metric for logistic regression in order to predict the SOFAS logistic outcome measure(it's a binary classification: 1 or 0). Lowest misclassification error on validation set: 0.3379138 and on test set we get misclassification error of .41.
Sensitivity and specificity of the ROC (Receiver Operating Characteristic) curve and AUC (Area under the curve) are used to understand the model performance for logistic regression.
In the picture below,
left hand side contains the ROC curve with AUC of 67 percent for logistic regression without excluding any of the feature variables and the right hand side contains the RMSE values for different regression

that summarizes the different techniques of supervised learning.

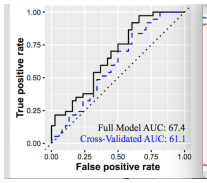
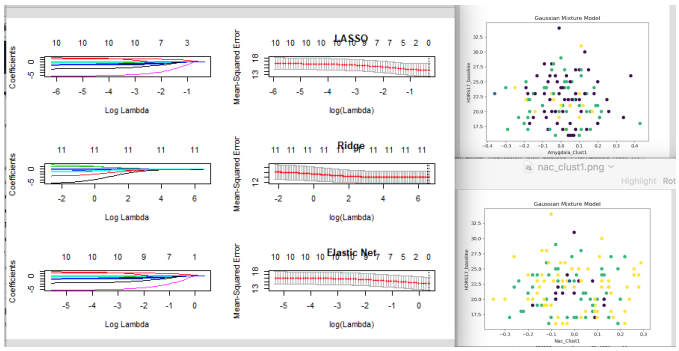


Table 1: RMSE for training and test data

Linear Regression : training= 3.4231	test= 3.6586
Ridge Regression : training=3.4244	test=3.6300
Lasso : training=3.5453	test=3.6862
Elastic Net : training= 3.5195	test=3.7181

Based on the RMSE values and the plots above for Supervised learning, Ridge regression performs the best. Hence, HDRS and SOFAS scores statistically connect antidepressants to Brain scan data.

To get more insight, we fit **Gaussian mixture model** using HDRS score and Amygdala Clus 1/2 brain data as well as HDRS score and Nac Clus 1/2 brain data, however based on the plot below, the representation seems unintelligible and requires further analysis.



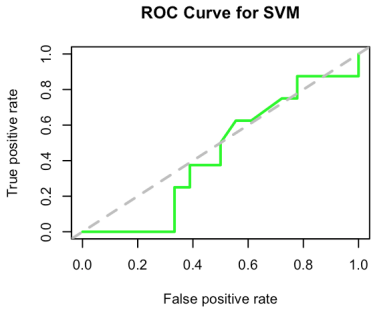
Factor Analysis output

In Factor Analysis, we transform the current set of variables into an equal number of variables such that each new variable is a combination of the current ones through some transformation. Here data gets transformed in the direction of each eigenvector and represent all the new variables or factors using the eigenvalues. An eigenvalue more than 1 means that the new factor explains more variance than the original variable. Output of our Factor Loadings shows that all 11 feature variables(3 antidepressants,age,gender,education,5 brain scan attributes) adequately represent the factor categories for this medical data set.

SVM classifier

is tested using three different kernels.
Here are the test errors for different types of kernels.
$Linear\ kernel$: 0.3745
$Radial\ kernel$: 0.34125
$Polynomial\ kernel\ of\ degree2$: 0.37825

Usage of Kernel depends on the data set. The linear kernel works best if the dataset is linearly separable, but if there is non-linearity then radial or polynomial kernel will produce better results. Radial kernel worked the best among all three kernels. This might seem obvious,because it is very likely to expect non-linearity among HDRS/SOFAS, all social attributes, like age,gender,education and 5 brain attributes in higher dimensions.
AUC for Radial kernel:0.4840278



VII. *Future Enhancements*

We would like to enhance our Gaussian Mixture Model with regression and sparsity[4] as follows: instead of estimating the $\mu_k\ for\ k = 1..K$, we would estimate only the coefficients of a sparse linear combinations of the X_is for all the data belonging to the same cluster using a sparsity enforcing penalty like l_1 norm of the coefficients. The main difficulty with such an approach might be to choose the right sample vector representing each cluster a priori, we would like to use Lasso[1] as one of the potential approach to solve that problem.[1]

VIII. *Github link*

The following github repo contains a link of the code and a copy of iSPOT-D dataset obtained from Dr.Adina Fischer,MD,PhD, a resident Stanford Psychiatry physician and a T32-funded postdoctoral fellow under the mentorship of Professor Leanne Williams and Professor Alan Schatzberg, Williams PanLab, Precision Psychiatry and Translational Neuroscience.
https://github.com/suvasis/cs229

REFERENCES

[1] R. Tibshirani, (1996) Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267288.

[2] SOFAS
https://kenniscentrum-kjp.nl/wp-content/uploads/2018/04/Social-Occupational-Functioning-Assessment-Scale-SOFAS.pdf

[3] G.J. McLachlan and D. Peel. (2000) Finite Mixture Models. Wiley

[4] A. Khalili and J. Chen, (2007) Variable Selection in Finite Mixture of Regression Models, Journal of the American Statistical Association, Volume 102, Number 479, pp. 1025-1038.

[5] Predicting Inpatient Discharge Prioritization With Electronic Health Records: arXiv:1812.00371
Anand Avati, Stephen Pfohl, Chris Lin, Thao Nguyen, Meng Zhang, Philip Hwang, Jessica Wetstone, Kenneth Jung, Andrew Ng, Nigam H. Shah

[6] https://dcf.psychiatry.ufl.edu/files/2011/05/HAMILTON-DEPRESSION.pdf

[7] Countdown Regression: Sharp and Calibrated Survival Predictions : arXiv:1806.08324
Anand Avati, Tony Duan, Kenneth Jung, Nigam H. Shah, Andrew Ng