A Machine Learning Approach to Predicting Dementia

Sagada Peñano
sagadap@stanford.edu
Department of Computer Science, Stanford University

https://case.edu/med/neurology/NR/Dementia.htm

Abstract

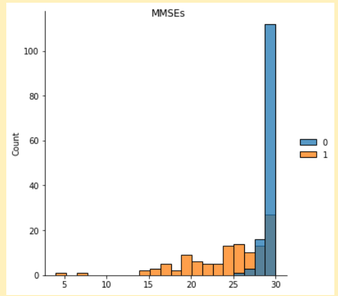
This work explores the use of machine learning algorithms on clinical data to predict dementia in patients. Factors that were taken into consideration were the patient's age, education level, socioeconomic status, mini mental state examination score, estimated total intracranial volume, and normalized whole brain volume, and atlas scaling factor. Multiple models such as logistic regression, random forest classifier, and decision tree classifier were tested with the random forest classifier producing the highest cross validation score of 85%. Furthermore, it was found that the most prominent features for determining dementia were mini mental state examination scores and normalized whole brain volumes.

Data

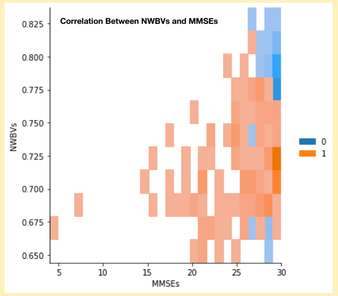
This project uses the OAS2 dataset from the Open Access Series of Imaging Studies (OASIS) . The dataset consists of 373 data points over 150 individuals that had 2-4 T1 weighted MRIs each. Patients with missing information were excluded, leaving a total of 354 examples as our final dataset. Each example consisted of 14 features, as well as a label of "Demented" or "Non-demented". The features listed were Subject ID, MRI ID, Visit Number, MR Delay, gender, dominant hand, age, level of education, social economic status, mini mental state examination score, clinical dementia rating, estimated total intracranial volume, normalized whole brain volume, and atlas scaling factor. To ensure that the algorithm was predicting solely on patient characteristics, Subject ID, MRI ID, Visit Number, and MR Delay were not considered. Furthermore, dominant hand was not considered because all patient's were right handed, and clinical dementia rating was not considered since it is assigned post diagnosis of dementia. This leaves us with eight features per example.

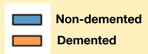
Features

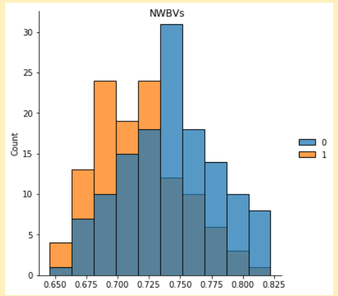
Feature	Feature Description
ASF	The Atlas Scaling Factor of a patient. This is the determinant.
Gender	The reported gender of the patient. Since all patients identified as male or female, this feature is represented as 0 for male and 1 for female.
Age	The age in years of the patient. This dataset contains data only from patients 60 or older.
EDUC	Years of education reported by patient.
SES	The social economic status of the patient, rated on a scale from 1 to 5. A score of 1 means that the patient belongs to the lower class of society, while a score of 5 means that the patient belongs to the upper class of society.
MMSE	The patient's score on the Mini Mental State Examination, which doctors administers to elderly patients to determine cognitive function. Scores can range from 0 (most likely to be demented) to 30 (least likely to be demented).
eTIV	The estimated intercranial volume of the patient in nm^3.
nWBV	Normalized whole brain volume of patient in mg.



MMSE







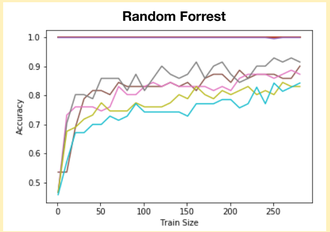
NWBV

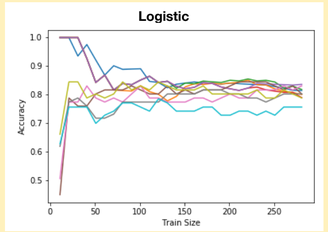
Models

To determine the optimal model, baseline tests were run over a number of different models using SciKit Learn, a machine learning library package for Python. Using SciKit Learn's default parameters, the following cross validation accuracies were achieved.

Baseline Classifier	Accuracy
Random Forrest Classifier	83.63%
Logistic Regression	79.65%
Decision Tree Classifier	77.68%
K Nearest Neighbors Classifier	67.8%
Neural Network	53.38%

An exhaustive search to find the best parameters for the top three models was then performed, and the models were retrained on their optimal parameters. With the optimal parameters computed by SciKit Learn, we find that the logistic regression and random forrest classifiers converge to roughly the same cross validation accuracy. However, it appears that the random forest classifier is overfitting while the logistic regression classifier seems to generalize well.





The random forest classifier creates multiple decision trees that each model the data in a slightly different way. To construct these decision trees, the Gini index is computed

$$Gini = 1 - \sum_{i=1}^{n} (P_i^2)$$

In this equation, Pi is the probability of a feature being classified for a distinct class. When constructing decision trees, features with a smaller Gini index would be preferred since they "split" the data more. Once the collection of trees is created, the final prediction is taken to be the average prediction of each individual tree.

The logistic regression classifier uses the sigmoid function as a hypothesis.

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

where theta is a parameter vector that assigns weights to features. At each iteration, the parameter is updated by

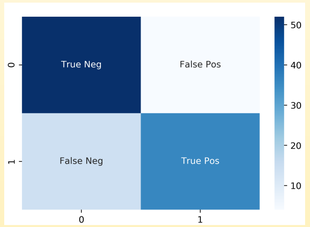
$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$$

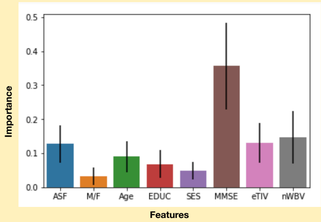
This update rule increments or decrements theta in the direction that maximizes the log likelihood of parameters or minimizes a specified loss. Once theta converges, the hypothesis function can be used on unseen data to make predictions.

Results

After optimizing hyper-parameters, both the logistic regression classifier and random forrest classifier had a cross validation accuracy of 84%.

The confusion matrices for both classifiers were very similar, with the majority of incorrect labelings being false negatives. This persisted among all train-test splits, indicating that the algorithm was biased towards the non-demented class.





Furthermore, the importance of each feature can be visualized from its Gini index. The bar plot to the left was constructed using the mean and standard deviation of cumulative impurity decrease within each tree.

We can clearly see that the most important feature is mini mental state examination score, followed by features dealing with brain size. Sociological factors such as education and socioeconomic status have little importance in predicting dementia.

Conclusion

A support vector machine classifier and logistic regression classifier were both able to achieve cross validation accuracies of 84%. Even though these algorithms had a decent accuracy, there is still much work to be done. The dataset that was used was fairly small with only 354 data points, causing the cross validation accuracy to fluctuate on the order of ~3% with each run. A larger dataset would greatly improve the model. Furthermore, now that we know that mini mental state examinations and brain characteristics are the most prominent factor in predicting dementia, we can try running the models on more features derived from MRIs.

Acknowledgements

I would like to thank the teaching staff of CS 229 for an amazing quarter and giving me the tools to put together this project. Data were provided in part by OASIS Longitudinal: Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. https://doi.org/10.1162/jocn.2009.21407.