Predicting Alzheimer And Cognitive Impairment From Handwriting Tasks

Martijn van Dam, Jonathan Koop and Merlin Urbanski
INFOMDA2

December 6, 2024

1 Introduction

1.1 Data

Neurodegenerative diseases, such as Alzheimer's, cause the brain and nerves to slowly stop working properly. This affects how people think, move and perform everyday tasks. There is no cure, but early detection of these diseases could help slow the progression of symptoms and improve life for those affected (Cilia et al., 2022).

One way to detect these diseases early on is by analyzing people's writing behavior (Cilia et al., 2022). Handwriting involves fine motor skills, suggesting that alterations in an individual's writing style could potentially provide insight into the condition of their neurological health. Despite many researchers having studied handwriting, there is no standard way to collect data. In addition to that, most studies are based on data with only a few participants. For our assignment, we will draw on the DARWIN (Diagnosis of Alzheimer with Handwriting) data set (Fontanella, 2022), comprising handwriting and drawing samples from a total of 174 participants. Out of these, 89 were Alzheimer patients while the remaining 85 were labeled as healthy, thus serving as a control group. The participants were asked to complete 25 different tasks (such as, retracing a circle, drawing their own signature) while 18 features (for instance, total time, paper time, air time) of each were measured. This leads to a dataframe with 174 observations and 450 features (18 for each of the 25 tasks).

1.2 How the original article dealt with the high dimensionality

The original article Cilia et al., 2022 addressed the challenge of high dimensionality by exploring two main approaches. First, they combined all the features from the 25 handwriting tasks into a single feature set for each task and used this to train a classifier. This allowed them to assess how well each task could distinguish between Alzheimer's patients and healthy individuals. Furthermore, it allowed them to assess which machine learning method worked best for each set of features.

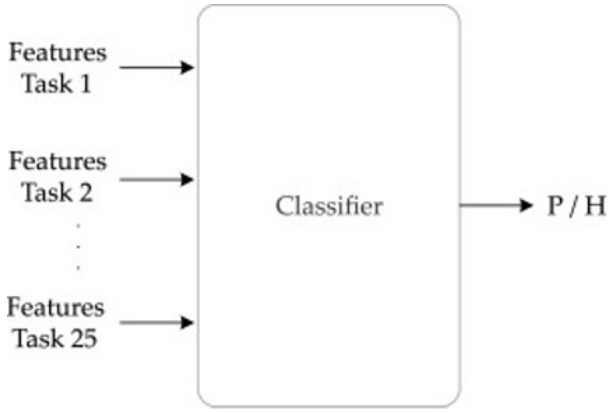

Figure 1: Representation of first approach for classifying individuals as Alzheimer/Cognitive Impairment Patients (P) or Healthy (H) (Cilia et al., 2022)

Second, they used the best performing model for each task and combined their outcome by majority votes to predict if the individual was a patient or healthy.

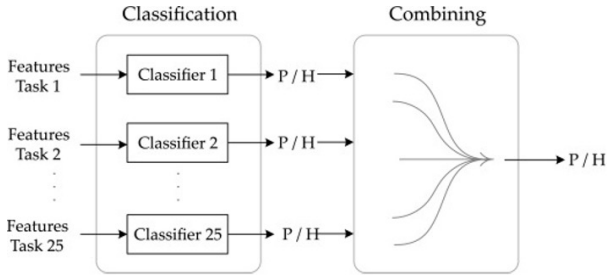

Figure 2: Representation of second approach for classifying individuals as Alzheimer/Cognitive Impairment Patients (P) or Healthy (H) (Cilia et al., 2022)

1.3 What we will do

In our assignment, we will evaluate the performance of logistic regression with principal components as predictors and compare it to the outcome of the original article, thus leading to the research question:

Does logistic regression analysis with principal components improve predictive performance in classification compared to a classical logistic regression approach?

In the Methods section, we will explain the procedure of the analyses and elaborate on how we compare the outcomes of the analyses. Next, we present the results of our analysis and provide a quantitative summary. An interpretation of the results, conclusions on the performances of the model, and suggestions for future research are described in the conclusion section.

2 Methods

To predict one's health status, we employ several approaches. The first analysis technique will employ a simple logistic regression model serving as a benchmark model. The choice of predictors used in this model will be rather ignorant: After applying t-tests for differences in means between patients and healthy controls, we will include those variables as predictors that show a statistically significant difference at the threshold of p ¡ .001. This is the threshold that reduces the number of predictors to a value smaller than n, allowing us to estimate an identifiable model. Furthermore, we are planning to replicate the analysis procedure and results by Cilia et al., 2022 as described in the previous subsection.

Following that, we will employ a logistic regression model with principal components retrieved from principal component analysis of all potential predictors. This procedure can be seen as appropriate, as all potential predictors are numeric. With PCA, we identify components summarizing the dimensions in which all potential predictors show the highest variance. The PCA transforms the original set of potentially correlated predictors into a smaller set of uncorrelated components, which reduces the complexity of the model. For this, the predictors need to be standardized, ensuring that all variables contribute equal to the analysis. The eigenvalue and eigenvectors of the covariance matrix are computed to identify the principal components. The principal components are then ranked based on the amount of variance they explain. Following this, we will compare the two selection criteria introduced in the course: First, inspecting a scree plot, we will apply the elbow criterion to determine the number of principal components to retain, selecting the components up to the point where the additional variance explained begin to level off. Second, we will draw upon the eigenvalue-greater-than-one rule, selecting components with eigenvalues (i.e. here, variances) above one. Given the assumption that the directions in which the predictors show the most variation are the directions that are associated with Y (James et al., 2013) and a linear relationship between the principal components and the outcome, we will then estimate a logistic regression model to classify the observations as patients and healthy controls. In the case of indications for non-linear relations of some of the principal components with the logit transformation of the dependent variable, we will consider including squared effects or differently transformed predictors and will otherwise resort to models not drawing upon the assumption of linearity.

To compare and validate our logistic regression models with the principal components, we employ a five-fold cross-validation with a test and train split and estimate the error rate as is done by (Cilia et al., 2022. We compare our principal component logistic regression models with each other and the benchmark logistic regression model on the estimated error rate, and the fact whether the models have converged to a stable solution. We argue that the converged model with the lowest estimated error rate is the best model for the data at hand. In addition, we will evaluate the specificity, sensitivity and positive and negative predictive values, since the data are aimed at supporting the diagnosis of neurodegenerative diseases (Cilia et al., 2022).

3 Results

Preliminary results and outline

- Without cross-validation, regression analysis does not converge. Classification leads to a train error rate of 0, indicating strong overfitting. With 5-fold-cross-validation we retrieve an error rate of .36 implying a poorly fitting model.
- Results of the PCA
- Interpretation of the results (quantitatively, i.e. which variables strongly influence which components)
- Results of the PC logistic regression with selected number of components (resulting from elbow and eigenvalue-greater-than-one)
- Comparison of confusion matrices on test data
- Short conclusion on which model performs best at prediction.

4 Conclusion

General outline:

- Summary: We have employed logistic regression with discriminant predictors, and with multiple principal components and one component.
- Discuss the fact that the logistic regression did not converge, so it does not suffice as the best model.
- Discuss results of the PC logistic regression analyses. Give substantive conclusions on which model is the best for our data.
- Interpret the PC in the context of the data.
- Dive into the meaning of the PCs in the context of our data: is it actually viable from a substantive point of view, or is it 'too data driven'? Potentially think about factor analyses approaches
- Additional limitations of PCA.
- Give suggestions for future review: if components are preferred, one can go for exploratory factor analysis and check whether theoretically sound constructs can be made.
- Overall (short) conclusion about this assignment.

References

Cilia, N. D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A., & Parziale, A. (2022). Diagnosing alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, *111*, 104822. https://doi.org/https://doi.org/10.1016/j.engappai.2022.104822

Fontanella, F. (2022). Darwin. https://doi.org/10.24432/C55D0K

James, G. M., Witten, D., Hastie, T. J., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. Springer.