**https://github.com/nikolas73/E-29-Graduate-Project**

Slide 1

My name is Nikolaos Giagtzoglou and I am a neurobiologist. I will present my Graduate Project for E-29. My aim is to analyze a proteomics dataset to identify important genes/Proteins for Alzheimers.

Slide 2

The course was offered in Spring 2018 at Harvard Extension School. Many thanks to Nenad and his team of teaching stuff for theiur efforts and support throughout the semester.

Slide 3

Alzheimer’s Disease (AD) is a devastative neurodegenerative disease and the leading cause of dementia. The burden of AD translates into billions of dollars annualy and is projected to increase along with the increase of aging population. Treating AD is difficult because of its complicated nature. To implement effective therapies we need to understand the molecular basis of the onset and progression.

Such understanding will rely heavily on the analysis of genomic data that have been collected throughout the world. I am planning to analyze one such public dataset to identify genes that predict Dementia. These can then be used as biomarkers or therapeutic targets.

Slide 4

The required libraries are shown here

Slide 5

The references are mainly from two O’riley books that have helped me understand certain aspects of logistic regression and the pipeline tool.

Slide 6

I thank Nenad, his teaching staff and my wife Isabel for her patience and support.

Jupyter Notebook

In the beginning I imported synapse client package that allowed me to login in the corresponding website of AMP-AD to download the proteomics file that I used for analysis.

Then I imported all necessary libraries/

First I loaded the protein data in a dataframe and then I loaded the disease traits ans sample identities in another.

Then I cleaned the data. In the protein dataframe, I either filled in the non existing values with a value of 1 - that helps in the log transformation later on – (PrIn4) or I just dropped them altogether (PrIn5). I then merged those two different dataframes with three columns of the trais dataframe, indicating Clinical dementia rating (CDR), mean size of Abeta plaques and sample identity.

In my exploratory analysis, Isaw by linear regression that there is a sorrelation between CDR and Plaque size (Figure 1). I reasoned that a continuous variable like size of amyloid plaques is going to be a great feature for modeling the protein abundance against via linear regression.

Protein abundance can vary a lot. So, one way to overcome this problem is to use the log10 transformation of protein abundance values, which brings them all at comparable scales (Figures 2 and 3).

Furthermore, one can employ a standard scaler. I wrote a function that examined which dataframe (PrIn4 or PrIn5) is best predicting in a linear regression model the size of amyloid plaques. It appeare that the best dataframe is PrIn5b, in which we have avoided imputation, after log10 engineering and standard scaling. However, the prediction was only ~30%, independently if we model against plaque size or against Clinical dementia rating (CDR).

In any case, I used the coefficients of linear regression to determine important genes. I selected those, which lie in the upper or lower 2.5%. These are 73 genes and their coefficients are in Figure 4.

It would be interesting to see whether these genes are coexpressed. To identify coexpression clusters I used KMEANS to cluster genes that vary accordingly among different samples. An initial inertia analysis (Figure 5) revealed that 10 clusters would be more than sufficient. Surprisingly, the vast majority of “important” genes that cae out of the linear regression analysis are found in the same cluster suggesting that they are coregulated in disease, as some of them have been previously related to Alzheimer’s such as ApoE (the most prominent AD risk factor), Abeta (product of APP) , C4B (component of complement system), etc.

I thought that the linear regression prediction would improve if Ireduce the dimension of the dataset by principal component analysis. In Figure 6B, it is shown that 85% of total variation can be explained by two different components. However, as shown in Figure 6, these components do not show any correlation after a linear model is applied for CDr or Plaque size. Therefore, I decide to explore new models.

I turned my attention to support vector machines that are able to recognize hyperplanes of separation. Im addition I simplified the outcome by rendering it a binary choice of “nonAD” and ”AD” samples if their Clinical Dementia Rating (CDR) was less or equal to 2 (nonAD) or not (AD). Using the non imputed values in PrIn5, I found that SVM model can predict at a rate of approximately 57%, in boththe full and the PCA reduced dataset.

This finding encourage me to use the pipeline tool in combination to Gridsearch tool to explore different conditions of modeling, including SVM, Logistic Regression and Random Forest Classifier. The best score was achieved with a Logistic Regression model and Standard Scaler preprocessin in the full, not logged10 dataset.

A multinomial logistic regression had poor output and was dropped. I ended up using a logistic Regression model on the log10 transformed non imputed dataset without scaling. The overall output was approximately 77%.

The coefficients of the genes that are 3 standard deviations away from the means of all the coefficients are shown in Figure 7. The total of these strictly chosen genes are 15, include Abeta (the product of APP processing and main constituent of amylid plaques in AD), may somewhat reliably predict the AD from non AD status. They can be useful biomarkers and therapeutic targets.

Thanks to Nenad and his team for a great semester.

Nikolaos Giagtzoglou