BMI 704: Data Science 1 Spring 2019

Short Assignment 3: Predicting age with molecular phenotypes

Date Due: 3/15/19 by 11:59pm EST.

You may work together, but each student must hand in their code and answers separately. This assignment is a total of 100 points (and 20 points/day will be docked for each day late). You can complete this assignment on O2, your own laptop, or the RStudio Server.

Submit your answers (e.g., figures) and as a .Rmd knitted file as a .pdf or .html file to Canvas.

Background:

Aging is a hot subject in biomedicine and the greater community. For example, perhaps you have tried to use Microsoft's age predictor "How old do I look" (https://how-old.net/), an attempt to predict your age based on an image of you.

Prediction of age can be important to understand the difference between *chronological* and *biological* age (see Background question below). Steve Horvath, a professor of bioinformatics at UCLA, built an age predictor using *DNA methylation* marks as predictors to examine potential differences in chronological and biological aging. In this paper, Prof. Horvath assembled a large set of publicly available DNA methylation (Illumina 450K and other array assays) data and asked a simple question: can DNA methylation levels predict chronological age? And if so, are deviations from predicted age evidence for accelerated aging?

But does this matter for medical decision making? A older indicator used for clinical purposes that summarize the pathological and disease manifestations of aging includes the Charlson Comorbidity Index, invented by Mary Charlson and colleagues. The Charlson index has proved useful in aiding decision making, such as clinical study design (inclusion/exclusion criteria) and risk stratification.

In this short assignment, we will download two large methylation datasets and build our own chronological age predictor as a function of DNA methylated sites in blood tissue and compare this to other "age predictors" that are used (off and on) in medicine today, such as the **Charlson Comorbidity Score.**

Description of Data and Reading:

Please read Steve Horvath's paper (linked here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015143/). Also, please read Mary Charlson's paper (linked here: https://www.ncbi.nlm.nih.gov/pubmed/3558716)
(All readings can be found on Canvas)

We have already downloaded the data from the Gene Expression Omnibus for you and they are located in **O2**:

1) **Training data (GSE40279):**
   `/n/groups/bmi704_spring18/assignment3/GSE40279_r2.Rdata`

2) **Testing data (GSE41169):**
`/n/groups/bmi704_spring18/assignment3/GSE41169_r2.Rdata`

Or if you prefer, you can access the load the data right on the **RStudio Server:**

1.) **Training data (GSE40279):** `/home/data/assignment3/GSE40279_r2.Rdata`
2.) **Testing data (GSE41169):** `/home/data/assignment3/GSE41169_r2.Rdata`

Each of these .Rdata files contains the following (e.g., GSE40279):

1.) A data.frame called 'gse40279.meta', the characteristics of each sample, including age of the participant. This data.frame is N number of rows, where N corresponds to the number of samples in the dataset.
2.) A data matrix called 'gse40279.data', a data matrix that is ~470K by N, where N is the number of participants.
3.) An expression dataset called 'gse40279.exprSet'. This is the raw data from which (1) and (2) are derived. This is not used, but included for completeness.

Questions (**100** points total plus **30** Bonus points)

Background Questions:

1.) What are/is the diseases of interest ascertained in Horvath's study? Draw the **E, G, P, and D** associative diagram (from Lecture 1 and 2) and annotate what is investigated in Horvath's study. (3 points)

2.) What is the biological role of "DNA methylation"? How and where is DNA methylated? (2 points)

3.) What is the assay platform for this study? Briefly describe the concept behind the assay technology (2 points).

4.) What statistical method did Horvath use? What R package encodes this method? Download and install this package onto your R installation. Write down the function call (assume the dependent variable, or variable to be predicted is called age and the methylation indicators are in a variable called x. (2 points)

5.) Both GWAS and predictive methods can be used for variable selection. Aside from the question and data modalities, what are the key analytic differences between GWAS, EWAS, and the method applied in Horvath's paper? (2 points)

6.) What is the biological relevance of "age acceleration" or "age deceleration"? How is this quantity estimated? (2 points)

7.) What is the sample size of the training (**GSE40279)** and the testing datasets (**GSE41169)? (2 points).** (2 points)

Predicting Age:

1.) Where is the age information in the data.frame(s)? Write code to get the age from the data.frame in both the training and testing data. Second, how many probes do they have in common? Create a new

training and testing dataset that only contains the methylation probes that are in common in both of the datasets. (10 points)

2.) What are the quartiles (25th percentile, 50th percentile, and 75th percentile, maximum, and minimum) of the distribution of age in each of the datasets? What is the mean and standard deviation? What is the sample size of the datasets? (5 points)

3.) The testing and/or training datasets may have missing values (e.g., NA). To avoid "throwing data away", we will impute missing data with the mean. Write code to find the mean value of each probe in the dataset and impute the missing value for a probe by the probe mean. For example, if patient 10 had a missing value for probe cg1234, then replace the NA for patient 10 at cg1234 with the mean of cg1234 from the non-missing values of the population. (5 points)

4.) Predict age in the training dataset by optimizing the lambda parameter by 10-fold cross validation using only the probes that are in common in **both the training and testing datasets using ElasticNet regression (alpha parameter is 0.5)**. What is the lambda? Output the final model coefficients in a csv file (without headers) where the first column is the DNA methylation ID and the second column is the coefficient. (20 points)

5.) What is the R^2 and correlation of predicted versus actual age of the model in the testing and training datasets (10 points)?

6.) Plot the predicted versus actual age of both the training and testing data. (3 points).

7.) In the testing data set, how many individuals are "Age accelerated"? "Decelerated"? (2 points)

Implementation of a "Biological Age" Charlson Comorbidity Index

1.) What statistical technique did Charlson et al use to develop their comorbidity index? Where did they sample their cohorts? (4 points)
2.) Implement a Charlson Comorbidity Index R function for patient. Specifically, the function should input the age and disease condition a patient (or a array of patients) has conditions outputs the comorbidity index. (3 point)
3.) You are the research administrator of a hospital biobank and one of your goals is screening consented individuals (aged 20 and over) for entry into a longitudinal study to assess long-term chronic disease risk. To do so, you need to screen for participants who have, at study entry, **low predicted risk** for death (according to the Charlson Comorbidity Index). A colleague approaches you to evaluate using Horvath's epigenetic age predictor to complement a potential participant's chronological age. To evaluate this potentially high-cost screening tool (vs. the low cost of the Charlson), answer the following questions:
    a.) As you know from Question 2, the Charlson Index inputs age (ie, each decade over 40 adds one point to the index). How many individuals in the *training* and *testing* cohorts had a predicted age that would increment or decrement their Charlson index? (6 points)
    b.) If there are individuals whose Index would be incremented or decremented, how much are they changed? What diseases or conditions would this increment or decrement be equivalent to having, according to Charlson Index? If no individuals are found to increment or decrement the index in (a), speculate on reasons why (6 points).

c.) Based on your answers in (a) and (b), what is your recommendation to your colleague about implementation of the epigenetic age predictor for this use case? (5 points)

4.) Suppose we measured the epigenetic age of all 604 patient participants in Charlson's study. Write down the Cox proportional Hazards function invocation in R to test whether epigenetic age and chronic conditions are associated with overall survival. (6 points)

Molecular Aging Bonus:

1.) How can one analytically test if age acceleration is statistically different in males versus females? Execute this test in the training and testing datasets. (2 points)

2.) Rank the coefficients in order from largest in absolute value to smallest. What genes are implicated by the methylated sites picked by your model? Hint: attaining the 450K annotation file from an additional R library (IlluminaHumanMethylation450k.db) may be useful for answering this question.(5 points)

3.) Execute "principal components analysis" on the training dataset with the probes that were selected by ElasticNet regression. First, compute the correlation matrix between the selected probes and input the correlation matrix into the builtin principal components method in R, called `prcomp`. How many principal components describe 90% of variation in the dataset? What does this imply? (5 points).

4.) Use hierarchical clustering to cluster the probes that were selected by ElasticNet regression. Choose the top 3 probes that glmnet found by absolute value of coefficient size. What other probes are "clustered with" these probes? Describe some of the genes that are in this cluster. Do they make biological sense? (5 points).