FA2016-LIS542LE Midterm -- Replicating a study.

Due Tuesday November 1, 2016 1PM Central Time.

The data directory for this class has a file called Arrowsmith.xls which contains a "gold standards" dataset used to model *relevance* in an approach to literature-based discovery (http://dx.doi.org/10.1093/bioinformatics/btm161). Your goal is to replicate the model described in Table S2 (in the supplemental data file) in that paper, write up an account of your investigation, and reflect on aspects that made the process easy or hard. You may use any resource available to you except solicit advice from another human being. As usual, upload a pdf file with your narrative answers as well as your R code and data as separate files with the appropriate extensions.

Suggested steps include:

1. Read the paper and learn about literature-based discovery.

2. Convert the Excel file (.xls) to comma-separated-value plain text file (.csv).

3. Load the csv file into R and construct the following attributes

X1 = 1 if (nA > 1 or A-lit size < 1000) and (nC > 1 or C-lit size < 1000), 0 otherwise
X2 = 1 if nof MeSH > 0 and < 99999, 0.5 if nof MeSH = 99999, 0 otherwise
X3 = 1 if nof semantic categories > 0, 0 otherwise
X4 = cohesion score if cohesion score < 0.3, 0.3 otherwise
X5 = -|log10(n in MEDLINE) – 3|
X6 = max(min(1st year in MEDLINE,2005),1950)
X7 = min(8,-log10(pAC+0.000000001))
I1 = 1 if Arrowsmith search = 'retinal detachment', 0 otherwise
I2 = 1 if Arrowsmith search = 'NO and mitochondria vs PSD'
Similarly for I3 through I6.
Y = 1 if target = 0 or 2, 0 otherwise

3. Get to know the dataset: assess the summary statistics, histograms, and pairwise scatter plots before and after your transformation. Are there missing values or outliers? Include screen shots as figures in your write-up.

4. Fit a logistic regression model and assess the validity of its assumptions and statistical significance. Interpret the parameters and your model. Are your parameter estimates different from the ones reported? If so, why?

5. Reflect on aspects that made the process easy or hard.