TCSS 588 Hands-on assignment #1, due Wednesday 4/11/2018 @ 10:15am

Credit: The data and ideas behind these exercises and homeworks are from the NIH LINCS DCIC Crowdsourcing Portal and Ma'ayan Lab @ Mt Sinai, New York.
http://www.maayanlab.net/crowdsourcing/megatask1.php

The overarching goal is to predict adverse drug reactions. This assignment builds on the in-class examples on the ADR (adverse drug effect) prediction. You will see this dataset again in future assignments, and will build on what you have done in this assignment.

This assignment will be graded out of 10 points, with optional bonus 2 points.
You can work in groups of 1-3 students for this assignment. Groups are strongly encouraged.

Upload the following files for this assignment:

Q1 and Q2: A Jupyter notebook file named *hw1.ipynb* containing the R code and answers for Q1 and Q2.
Bonus question Q3: Since the bonus question is computationally intensive, you can choose to submit either
a notebook file named *hw1bonus.ipynb* OR
one **single** R script named "*hw1bonus.R*" containing all the code + a document file named "*hw1bonus.pdf*" containing the written answers.

In your R code, you can assume the files "**gene_expression_n438x978.txt**" and "**ADRs_HLGT_n438x232.txt**" are in your working directory.

Note that Q2 and Q3 will take substantial computational time to finish. Do not leave this until the last minute.

(Total 5 points) Correlation and data exploration

1. Using the following data "**gene_expression_n438x978.txt**" to answer this question.
   a. (1 point) Write R code to compute the correlation coefficient between all pairs of drugs. Then, plot a histogram of the correlation. What is the median correlation across all pairs of drugs?

   b. (1 point) Name the drug pair that gives the highest correlation coefficient (i.e. closest to 1). Produce a scatter plot to show the relationship between this pair of drugs using R.

   c. (1 point) Name 10 drug pairs that give the top 10 highest correlation coefficients (i.e. closest to 1).

   d. (1 point) Name the drug pair that gives the lowest correlation (i.e. closest to -1). Produce a scatter plot to show the relationship between this pair of drugs using R.

   e. (1 point) Name the most similar drug for each of the following:

i. CLOFARABINE
ii. DAUNORUBICIN
iii. FLUDARABINE

(Total 5 points) Logistic regression and model selection

2. Using the following data "**gene_expression_n438x978.txt**" and "**ADRs_HLGT_n438x232.txt**" to answer this question.

   In class, we used forward stepwise logistic regression to build predictive models for side effect "heart failure" by considering only the first 20 genes (variables) in "**gene_expression_n438x978.txt**" and using *all 438* drugs.

   Answer this question using all 438 drugs.

   a. (2.5 points) Start with the NULL model. Consider the first 50 genes (columns 1:50 in "**gene_expression_n438x978.txt**"), use forward stepwise logistic regression to build predictive models for each of the 232 side effects in "**ADRs_HLGT_n438x232.txt**". Which side effect can be predicted with the smallest number of errors? What is the corresponding model?
   Which side effect can be predicted with the smallest AIC? What is the corresponding model?

   b. (2.5 points) Repeat Question (2a) using backward elimination instead of forward stepwise. Start with the FULL model. Consider the first 50 genes (columns 1:50 in "**gene_expression_n438x978.txt**"), use backward elimination in logistic regression to build predictive models for each of the 232 side effects in "**ADRs_HLGT_n438x232.txt**".
   Which side effect can be predicted with the smallest number of errors? What is the corresponding model?
   Which side effect can be predicted with the smallest AIC? What is the corresponding model?

(Bonus 2 points) Model selection and cross validation
Note: this is a *very* computationally intensive question. If you want to attempt this bonus question, you need to get started way ahead of the due date.

3. We also saw in class that using the same training and test sets can lead to over-fitting. In this question, answer questions (2a) and (2b) using 10-fold cross validation instead of the entire training data consisting of 438 drugs. Due to the random nature of cross validation, repeat this process 3 times.

   Due to the time consuming nature of this question, you can use the first 20 genes in Q3 instead of the first 50 genes (as in Q1 and Q2).

Report the average number of errors over 3 runs of cross validation. Don't worry about returning the models from each fold and each run since the models would change when the training data change.

If you run out of time and can only finish 1 or 2 cross validation runs. Turn in what you have, document and you will receive partial credits.