#homework (3 questions, A-C. Make sure to expand all bullets!)

- A: Complete exercise 2 from https://stephens999.github.io/fiveMinuteStats/likelihood_ratio_simple_models.html
- B: Consider the zipcode data from ESL (https://web.stanford.edu/~hastie/ElemStatLearn/index.html)
  - Download the data and try plotting a few examples of the training data as 16 x 16 images to see if you can see the digits visually as expected.
  - Consider the problem of trying to distinguish the digit 2 from the digit 3.
    - Use the training data to learn classifiers, using a) logistic regression and b) k-nn, with k=1,3,5,7,15. This gives 6 classifiers in total. Misclassification rate is actually 0-1 loss
    - Apply these classifiers to the test data, and plot the misclassification rates for both training data and test data. (Plot the results for k-nn with k on x axis, and misclassification rate on y axis, with two different colors for test and training sets. Then put appropriately colored horizontal lines on the same plot -- one for test and one for train -- indicating the results for logistic regression.)
    - (Note: This is basically Exercise 2.8 from ESL, except replacing linear regression with logistic regression.)
  - Repeat the k-nn training as above, but using CV *on the training set* to tune k. That is, act like you do not have access to the test data and have to decide what k to use. How does it do? randomly select the data
  - Suppose now that for some reason it is considered worse to misclassify a 2 as a 3 than vice versa. Specifically, suppose you lose 5 points every time you misclassify a 2 as a 3, but 1 point every time you misclassify a 3 as a 2.
    - Modify your logistic regression classifier to take account of this new loss function. Compute the new loss on the test set for both the modified classifier and the original logistic classifier. train calssifier: weighted log-loss (log-likelihood) test report: weighted 0-1 loss
    - As far as you can, repeat this for the k-nn classifiers (ie modify them for the new loss function and compare the loss for modified vs original classifiers). Discuss any challenges you face here.
- C: Clone the repository https://github.com/stephens999/stat302/.
  - Run the R code in `exercises/seeb/train_test.R` (ie https://github.com/stephens999/stat302/blob/master/exercises/seeb/train_test.R) which loads and processes a dataset on 544 fish (salmon) at 12 genetic markers/loci, from \cite{seeb:2007}.

- Complete the exercise in the commented text at the end of this file. (note: see https://github.com/stephens999/stat302/tree/master/data/seeb for a description of the data, and particular explanation of the idea that fish are "diploid").

```
code & some examples
```