# STA 4364 HW 4

**Submission Format: Please submit your homework as 1) a HTML or pdf document, and 2) also submit the source file in either R Markdown or Jupyter notebook format (at most one of each type of file)**.

Problems can be done in Python or R. ISL = Introduction to Statistical Learning textbook.

**Problem 1** In this problem, you will predict tumor type from gene expression data. Since there are many more gene features than observations of patients, we will use ridge and LASSO regularization for logistic regression to reduce overfitting and help select the most relevant features out of a large group of features. This dataset has a multi-class outcome variable. The possible tumor types are BRCA, COAD, KIRC, LUAD, or PRAD. You will analyze this dataset by building a multinomial regression model with $\ell_1$ and $\ell_2$ regularization. The recommended approach is the `glmnet` package in R, which is covered in the code in class. You can check the "Multinomial Regression" section found at this link for specific information about multinomial regression in `glmnet`.

**Note:** This is quite a large dataset so models will take a minute or two to fit.

(a) Load the labels and data with `read.csv`. Remove any columns with missing entries. Remove any columns with variance less than 0.001. Standardize each gene predictor column to have mean 0 and standard deviation 1 (this is important when doing regularized regression). Split the dataset randomly into a training and validation set.

(b) Use ridge logistic regression with 10-fold cross validation to model the response given the gene expression predictors. What is your optimal value of the regularization parameter $\lambda$? Apply your model to give predictions using the optimal value of $\lambda$. Make a confusion matrix showing the accuracy of your model on the training and test set.

(c) Use LASSO logistic regression with 10-fold cross validation to model the response given the gene expression predictors. What is your optimal value of the regularization parameter $\lambda$? Apply your model to give predictions using the optimal value of $\lambda$. Make a confusion matrix showing the accuracy of your model on the training and test set.

(d) Give a list of the top 20 most relevant genes that are selected by your LASSO model at the optimal value of $\lambda$. The coefficients for a multinomial regression model will be a $p \times C$ matrix where $C$ is the number of classes and $p$ is the number of feature columns. What relation do your selected genes have to tumor expression? You can determine this by looking at which of the $C$ coefficients associated with a certain gene are non-zero. Positive values in a certain index correspond to a high probability of the tumor associated with that index, while negative values correspond to a lower probability.

**Note**: If columns are highly correlated, LASSO will often arbitrarily select a single column, so a full report of relevent genes would involve predictors selected by LASSO and genes that are highly correlated. Other techniques like group LASSO can select subsets of related genes, these will not be covered in this class. You could also try to combine the $\ell_1$ and $\ell_2$ penalties to get representation of meaningful predictors that are also correlated (see the reference material above).