Homework 1:

Due date: **Friday October 6 (midnight)**
Please email me your response (in a PDF format), your code and all necessary material in one compressed file (e.g., ZIP). A 10% penalty per day will be applied to any late submission (see the course outline for details).

In the following problems, we are going to use transcriptomic data (expression of thousands of genes) corresponding to hundreds of cell lines. From a machine learning (ML) perspective, cell lines correspond to samples and genes correspond to features.

You can access the gene expression data from the following link:
https://www.dropbox.com/s/vwhtcf7rdko26tw/TG_LASSO_GeneExpressionInput.zip?dl=0

Use the file titled "gdsc_expr_postCB.csv". In this file rows (with names starting with "ENSG") correspond to genes (treat them as features) and columns correspond to different cell lines (treat them as samples/instances/examples). Note that cell lines are represented using numbers.

In all questions below, unless otherwise specified, you can **choose** the parameters and options of each algorithm. However, you need to clearly and completely specify your choices.

1- Dimensionality reduction (30pts):
   a. Let's start by visualizing the data using dimensionality reduction methods. Use PCA, UMAP and t-SNE algorithms to map the samples (i.e., Cell Lines) onto a 2-D space for visualization (you do not need to implement these methods from scratch. You only need to find their implementation).
   b. Do you see different clusters forming (visually)? How much correspondence exist between clusters that you find in different methods?
   c. What is the difference between the three methods above? Use pros/cons to compare them against each other.

2- Clustering (30 pts):
   a. Use the dataset above and cluster cell lines using agglomerative clustering and k-means (default parameters) and identify 3 clusters using each method.
   b. How much concordance do you see between the clusters obtained using these methods? Use Jaccard Similarity to measure the consistency between pairs of clusters and form a table of 3x3 showing these values (rows can correspond to clusters formed using k-means and columns using agglomerative).
      https://scikit-learn.org/stable/modules/model_evaluation.html#jaccard-similarity-score
   c. Use Rand-Index and Adjusted Rand Index to determine how consistent the two algorithms are with respect to the identified clusters.
      https://scikit-learn.org/stable/modules/clustering.html#adjusted-rand-score

d. Use agglomerative clustering with "average" linkage and try distance (affinity) measures Euclidean and cosine to cluster samples into 3 groups similar to part a. Repeat the exercise in part b and form a 3x3 table to assess the consistency between clusters formed using these two distances (use Jaccard similarity). What can you conclude? Additionally, use the Rand-Index and Adjusted Rand-Index similar to part c. What do you observe?

3- Regression (40):
Now, let's try some simple regression algorithms. Here, for the feature matrix X, we use the gene expression data from the link "https://www.dropbox.com/s/vwhtcf7rdko26tw/TG_LASSO_GeneExpressionInput.zip?dl=0" and the file "gdsc_expr_postCB.csv".

For the response vector y, use the drug response values for drug Doxorubicin from the file "gdsc_dr.csv" across different samples from the link
"https://github.com/emad2/TG-LASSO/tree/master/Data"

Make sure you match the sample names (cell line names represented using numbers) in both files and remove any sample for which no response to Doxorubicin is recorded in gdsc_dr.csv.

a. If you recall, LASSO is a method based on linear regression that also has a L1 regularization term to impose sparsity. Consider the formulation below for LASSO, where X is the feature matrix, y is the feature vector (corresponding to Doxorubicin), \alpha is a hyperparameter that decides the strength of regularization term and w is the set of weights.

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha ||w||_1$$

Given this formulation, choose the value of \alpha to be [0.01, 0.1, 0.3, 0.5, 0.9]. Draw a graph that shows the number of selected features by LASSO vs. the value of hyperparameter.

b. Form a nested cross-validation (CV) to choose the best value of alpha and also to assess the performance of LASSO with the best hyperparameter in predicting the response to Doxorubicin. Use two measures Spearman rank correlation and mean squared error for the assessment. For the inner CV, use 3-fold CV and for the outer CV use 4-fold. Report the following:
    i. In details, describe how you performed the CV process
    ii. What was the best value for alpha
    iii. What was the average Spearman correlation / MSE between estimated values and true values?