Homework 2 (100 points, Due date: Monday, April 15th, 11:59 PM):

In this homework assignment, we will apply a supervised learning approach (kNN classifier) to predict drug sensitivity for a panel of cell lines based on gene expression profiles.

These data were taken from the 2012 NCI-DREAM Drug Sensitivity Prediction Challenge, which is an annual competition held in the computational biology community (read about Sub-challenge 1 details here: https://www.synapse.org/#!Synapse:syn2785778/wiki/)

Briefly, the cell lines were derived from a set of breast tumors or normal tissue. Each of the cell lines was exposed independently to five drugs, and the GI_{50} was measured. GI_{50} represents the drug concentration at which growth is inhibited by 50%. The GI_{50} values have been $-\log_{10}$ transformed so higher values reflect greater sensitivity. For simplicity, in this assignment we have binarized the drug sensitivity values such that "1" corresponds to sensitive and "0" corresponds to resistant.

Your goal is to build a kNN classifier that predicts which cell lines are sensitive to each drug.

Note: If you'd like to read how the best computational biology groups in the world approached this problem and how they performed, a paper was published following the competition describing the results of this challenge: Costello et al. A community effort to assess and improve drug sensitivity prediction algorithms Nature Biotechnology 32, 1202–1212 (2014).

Data files:

We have already merged the drug sensitivity and gene expression data into a single text file, called `DREAM_data.txt`. This is a tab-delimited file with the cell line IDs across the first row, and the drug sensitivity and gene expression data in the rows that follow. Please read the following important notes:

Drug sensitivity data (2nd to 6th rows of the file): the first column is the drug name and the rest of the columns are the drug sensitivity values (1=drug sensitive, 0=not sensitive). "NA" means the measurement is missing.

Gene expression data (7th row to the end): the first column is the gene name and the rest of the columns are gene-level summaries across breast cancer cell lines. Gene-level summaries of expression are already quantile normalized and log2-transformed.

Questions:

1. **k-NN implementation (40 points):** Implement a kNN classifier, using the Pearson correlation coefficient as the similarity metric between cell lines' expression profiles. You should create a classifier for each drug that takes a cell line expression profile as input and produces a score that predicts whether it is sensitive or resistant to the given drug.

2. **kNN performance evaluation (30 points):** Set k=5 for all of the evaluations in this problem. Apply leave-one-out cross-validation (LOOCV) to measure the performance of your classifier. For each of the 5 drugs, plot an ROC curve with sensitivity on the y-axis and (1-specificity) on the x-axis. Use the number of drug-sensitive neighbors (among the *k* total) for each cell line to rank the predictions in order to draw the ROC curve. For each ROC curve, be sure to plot the performance expected by a random classifier. Answer the following questions:
   (a) Does your classifier work better than a random classifier? For which drugs? Refer to specific evidence from your analysis to justify your answer.
   (b) For the drug on which your approach appears to work the best, how good is the classification performance? Pick a point on the ROC curve, and report the relevant metrics (number of true positives, false positives).

3. **Exploration of parameters affecting kNN performance (30 points):** For this problem, we will explore how the performance is influenced by the key parameter in the kNN classifier, *k*, and the scoring function.
   (a) Rerun your classification results with *k*=3,5,7. For each drug, create a single figure, but plot the ROC curves for all values of *k* on the same curve. Again, sort the predicted cell lines by the number of their nearest neighbors that are drug-sensitive. Discuss the results of your analysis. Does the choice of *k* affect the performance of the classifier?

   (b) Set *k*=5. Instead of scoring each cell line with the number of its nearest neighbors that are drug-sensitive, use the following weighted score:

$$S(x) = \sum_{i=1}^{k} sign(y_i) * PCC(x, y_i)$$

   Where y_i are the nearest neighbors and sign(y_i)=+1 for drug-sensitive cell lines and sign(y_i)=-1 for drug-resistant cell lines, and PCC is the Pearson correlation coefficient. Sort the predictions of drug-sensitivity by this score. Plot new ROC curves for each drug comparing the approach from Problem (2) with this approach.

Extra credit #1 (5 points): repeat #2 with at least one of the following classification models: SVM, decision tree, logistic regression, or neural networks.

Extra credit #2 (5 points): repeat #2 using the kNN approach but evaluate the impact of using the other genomic data associated with each cell line (e.g. RNAseq-based expression profiles, copy-number variation, proteomics data). These data can be found in `extra_data.zip`. Can you improve the classification performance by using these other data?

What to submit:
Please include all figures and analysis in a single PDF document. Put this document in a .zip or .tar.gz file, along with all source code you wrote to complete the assignment. Submit these files on the course Moodle site.