HW3: Programming Assignment

Predicting Drug Response in Tumor Cell Lines

Due Thursday October 24th at Midnight.

For each of these problems, use 5-fold cross validation and display the confusion matrix, f1 scores and other evaluation metrics as appropriate in addition to the accuracy or MAE achieved.

Please write a ½ page commentary on your assessment of your efforts on HW3.

Data for this problem includes the following:

1. Transcriptomes (RNAseq), and Genome Variation (SNPs) characterizing the cell lines from various studies (GDSC, CCLE, gCSI, CTRP, NCI60)
2. Combined "dose response" data, including drug, cell lines, dose, growth values.
3. Aggregated response data, drug, cell lines, computed AUC, etc.
4. Drug related data, Dragon7 descriptors, ECFP, PFP, SMILES, etc.
5. ALMANAC study including dose response for drug pairs

They can be found in the shared **MLiM-Datasets/HW3** directory.

I've precomputed data_frames for ~120 drugs from CTRP to be used in part one. Should choose ten of these to build models from. These files are in HW3/Part1 directory and labeled "**By.Drug.DRUG_ID.tsv**".

For Part2 you should use the "**df_top21_bal_AUC.tsv**" dataset which can be found in HW3/Part2.

For each part please turn in your code (a python notebook is a reasonable way, but a python script is also fine), turn in output from the program (text of graphics, graphics preferred). And a short (1 paragraph write up for each part and section, explaining what you did and your critique of the results, comments on problems or difficulties and possible future approaches that might do better).

Part 1. "By Drug" Tumor Response (50 points).

a. Using SciKit Learn build a machine learning regressor that predicts the AUC response for tumors for each of 10 drugs selected by you from the 250 drugs available. Sweep through multiple ML methods. Report on which drugs were best for which tumors and which ML methods performed the best.

b. **Extra Credit (20 points):** Using feature selection methods of your choice determine a < 100 gene (RNA-seq) signature that can be used to predict dose response for each of the

drugs. Determine how many genes in the compact signature are in common between your selected drugs.

c. Using **Keras**, build a deep learning classifier that performs the same regression task as part a for the responses to the 10 drugs you have selected. Compare the accuracy from part a to part c.

d. **Extra Credit (20 points):** Use the TPOT autoML system to search for a better solution to part a.

Part 2. **Dose independent formulation (50 points).**

a. Using SciKit Learn build a machine learning classifier that predicts drug response using the aggregated **df_top21_bal.tsv** dataset. Use AUC1 as the label. Threshold AUC1 (> 0.50 = 0, <= 0.50 = 1) for the prediction target to make it a classification problem. Input data includes tumor features and drug features.

b. Using **Keras**, build a deep learning classifier that performs the same classification task as in part a.

c. **Extra Credit (20 points)**: Explore the use of different features to represent the drugs, try drug descriptors, fingerprints or some other representation. Determine which types of features are more predictive.