Coding Lab #2 Predicting Drug Response in Cell Lines

The primary goal of this lab is to have you build one or more types of machine learning models for predicting drug response. We will use data from the GDSC. This data consists of drug response data from 265 drugs tested on ~1000 cell lines as discussed in class.

You can find the data at nciftp.cels.anl.gov

Login with user ID "anonymous" and send your email as the password.

You will need data from the GDSC directory.

I've copied and dumped data into CSV Files in this Dropbox Directory help you get started.

https://www.dropbox.com/sh/xbzlvn7l984ckb2/AABPf9-j2nLzFbvUDJ8zTJwaa?dl=0

There are also excel spreadsheet files in this directory that have more information about some of the data types, so you might want to look at those especially for section #2 of the lab.

The lab has three parts plus an extra credit option. :-)

1. Learn a model that can predict drug response on a per drug basis (by Drug) across the GDSC cell lines. You should use at least five fold cross validation (split the data 80% for training and 20% for testing, and repeat selecting disjoint five disjoint testing sets). Report the (Accuracy, RMSE) F1 score and AUC for these runs. You should use just the gene expression data for this part of the lab. You should experiment with predicting IC50 values as a regression problem and with a three way categorical classification [Sensitive, Intermediate, Resistant] (SIR) for the classification problem. For the SIR score you need to choose the thresholds. You can use any machine learning method you want for this or a combination of methods. Apply this modeling approach to all the drug response data that is available. Report your top 10 results and your bottom 10 results.

2. Try to improve the results of "by Drug" prediction by adding additional assay types to your model. In the data directory are datasets for copy number, WES variants, RACS in addition to the RMA normalized gene expression which you use in part 1. Try one or more of these to improve the model.

3. Develop a version of your model that can rank order the drugs for a given Cell Line. Report the top ten drugs predicted to be sensitive for each cell line in order of best to worst. Compare this list for each cell line to the experimental data in the GDSC experimental results. Come up with a score to compare your rank ordered list to the rank ordering of IC50 from the Dose Response data.

Good luck and report any problems or questions on Piazza.

4. Extra Credit. If you finish these and are still hungry. Consider the problem of adding some drug information to build a model that can predict for more than one drug. Compare the performance of this model to that in 1-3.

The assignment is due two weeks from Today (Due on Friday November 10th)