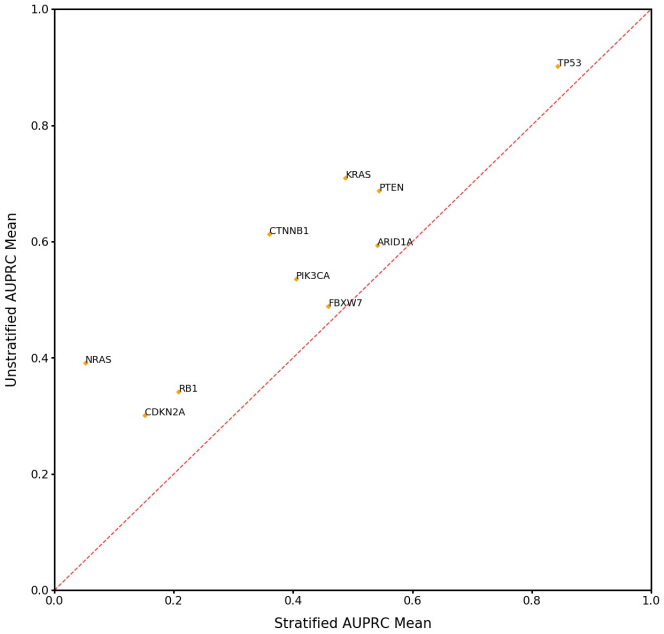
COBB2080: Directed Study 12/8/2020

Final Update

Name: Kamaru-Deen Lawal

Email: kal246@pitt.edu

At the beginning of the semester we set out to determine whether or not we would achieve better performance on the TCGA dataset by stratifying based on cancer type vs. random k-Fold cross validation. The main goal is to develop methods that identify functional mutations in targeted genes by the downstream methylation changes that they induce. Preliminary papers and results have shown certain regularization strategies produce interpretable, high-quality models. So we decided to go about attacking this problem by using an ElasticNet based classifier. Before getting into our final results we will do a quick summary of the dataset. The Cancer Genome Atlas (TCGA) is a cancer genomics program that has characterized 20,000 primary cancer and matched normal samples spanning 33 cancer types. The dataset is responsible for deepening our understanding of cancer through molecular characterizations and bolstering the computational biology field at large by allowing researchers to develop tools for a wide range of purposes. In order to assess whether or not stratifying based on cancer type vs. random k-Fold cross validation was better we tested out the procedure of performing stratified vs. unstratified cross validation for ten different targets. The specific proteins we used as targets were TP53, PIK3CA, KRAS, PTEN, ARID1A, RB1, FBXW7, NRAS, CDKN2A, and CTNNB1. For each of the proteins we averaged the AUPRC for each fold with the other folds for that protein to get a mean value to approximate the performance of the model for the given target. After doing this for both the stratified and unstratified methods we plotted the resultant mean AUPRCs against each other. The final results can be seen below.



Points lying directly on the diagonal ($y = x$) imply the stratified and unstratified methods performed the same, while points to the left and right of the diagonal imply the unstratified and stratified methods performed better, respectively, Given that all of the points lie on the left hand side of the diagonal we conclude that the unstratified performed better. There are multiple directions we could take to build on the work we began this semester. The most obvious direction would be to perform a more thorough

form of parameter optimization for both the number of folds in the unstratified for the case and for the parameters that are being used for the ElasticNet model.

Citations

- Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, Sander C, Cherniack AD, Mina M, Ciriello G, Schultz N; Cancer Genome Atlas Research Network, Sanchez Y, Greene CS. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep. 2018 Apr 3;23(1):172-180.e3. doi: 10.1016/j.celrep.2018.03.046. PMID: 29617658; PMCID: PMC5918694.
- F. Souza and R. Arajo, "Mixture of Elastic Net Experts and its Application to a Polymerization Batch Process," 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), Porto, 2018, pp. 939-944, doi: 10.1109/INDIN.2018.8472056.