Homework 3

ECS 171

Ryan Gosiaco 912819444

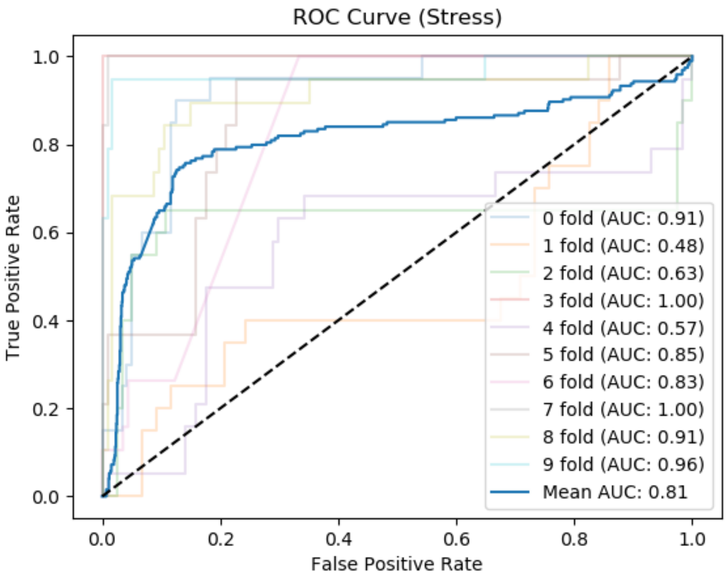
1. The regularized regression technique I used was Lasso. For this problem, I used LassoCV from sklearn and it optimizes

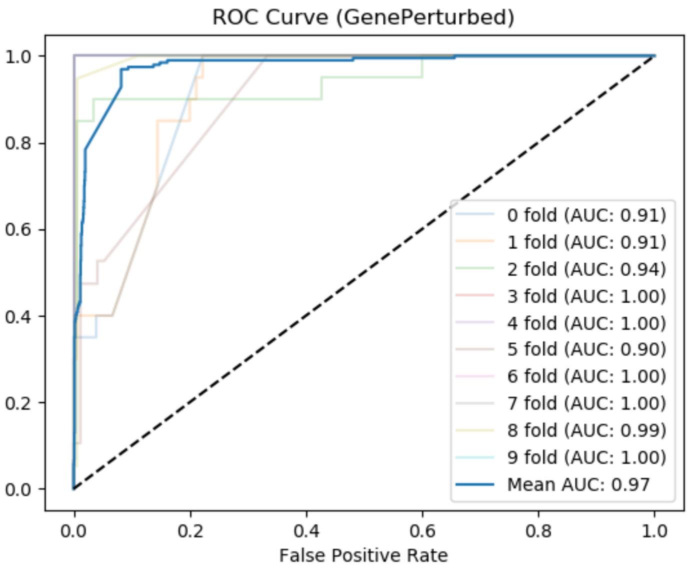
$$\left(\frac{1}{(2 \, * \, n\_samples\,)}\right) * \; ||y - Xw||_2^2 + \; alpha \; * \; ||w||_1$$

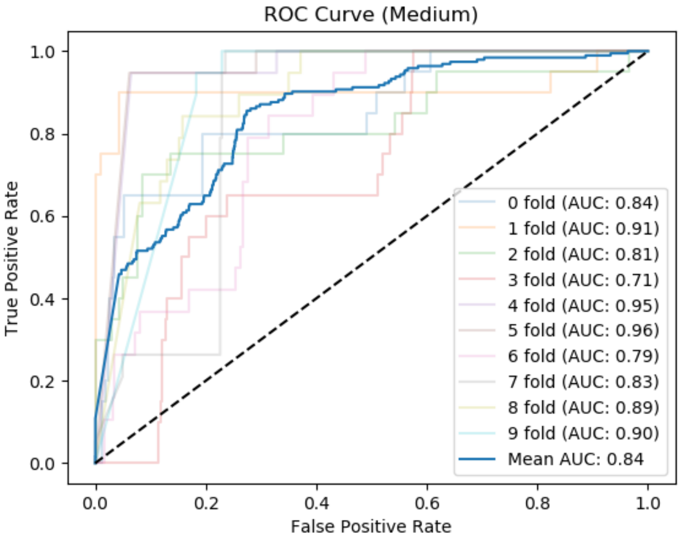
where alpha is the optimal constrained parameter value that we are looking for. The best alpha was 0.0004942. Through this, Lasso chooses some coefficients to go to zero which results in a simpler model. With a 10 fold cross validation, a maximum of 300 iterations and a tolerance of 0.01, Lasso results in 187 non-zero coefficients and has a generalization error of 0.03564.

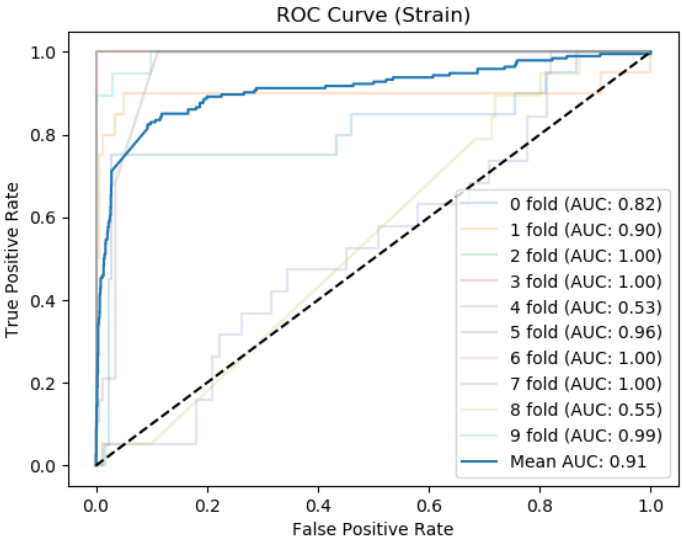
2&3. For the bootstrapping I did 100 iterations and for each iteration I sampled a training set the length of the dataset with replacement. I then trained the Lasso model with 1000 max iterations and a tolerance of 0.01. I then predicted the growth using the mean expression value and calculated the MSE between the prediction and the mean of the growth column. I am assuming that the true value of the mean expression value is the mean of the growth. With 95% confidence, the MSE is likely to be between 0.000 and 0.001.

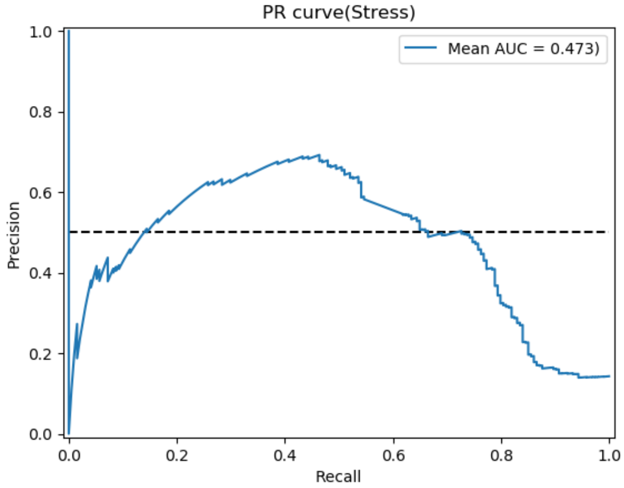
4. I used the feature selection method from Q1 which resulted in 187 non-zero features used. I used a OneVsRestClassifier and SVC with the rbf kernel, C equal to 1 and gamma set to scale. I also had to binarize the labels.

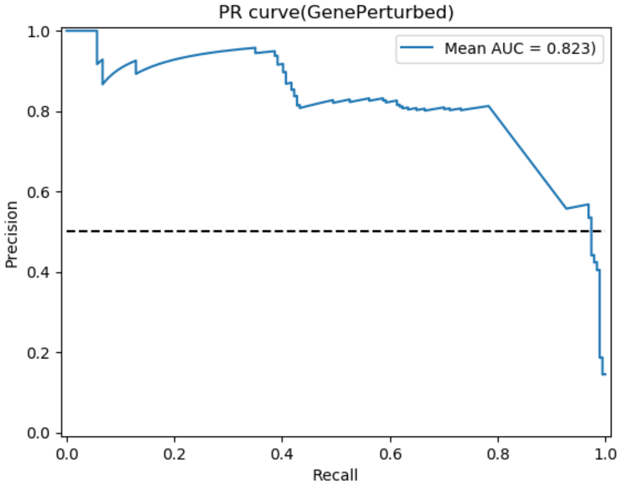


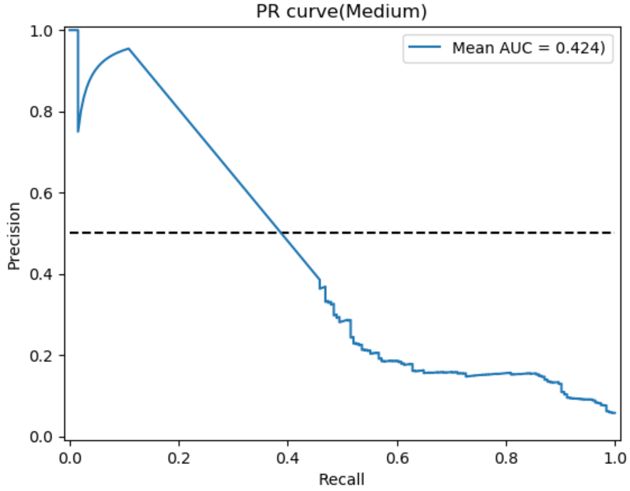


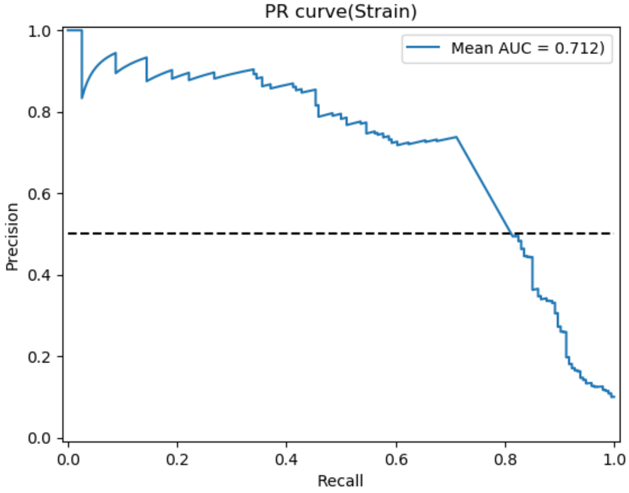

ROC Curve (Medium)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0 fold (AUC: 0.84)
1 fold (AUC: 0.91)
2 fold (AUC: 0.81)
3 fold (AUC: 0.71)
4 fold (AUC: 0.95)
5 fold (AUC: 0.96)
6 fold (AUC: 0.79)
7 fold (AUC: 0.83)
8 fold (AUC: 0.89)
9 fold (AUC: 0.90)
Mean AUC: 0.84


ROC Curve (Strain)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0 fold (AUC: 0.82)
1 fold (AUC: 0.90)
2 fold (AUC: 1.00)
3 fold (AUC: 1.00)
4 fold (AUC: 0.53)
5 fold (AUC: 0.96)
6 fold (AUC: 1.00)
7 fold (AUC: 1.00)
8 fold (AUC: 0.55)
9 fold (AUC: 0.99)
Mean AUC: 0.91


PR curve(Stress)
Mean AUC = 0.473)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

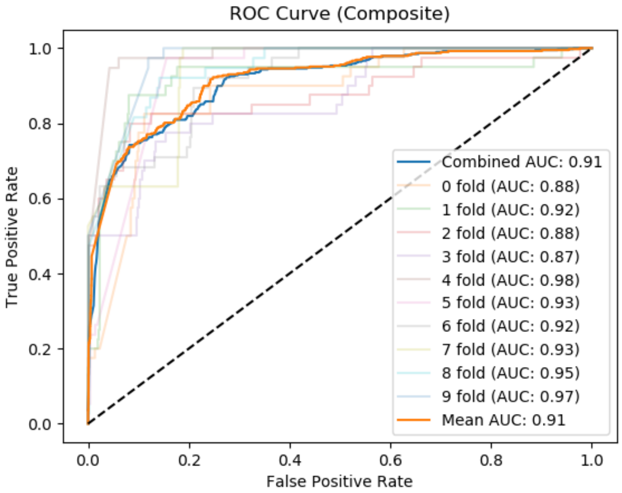

PR curve(GenePerturbed)
Mean AUC = 0.823)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

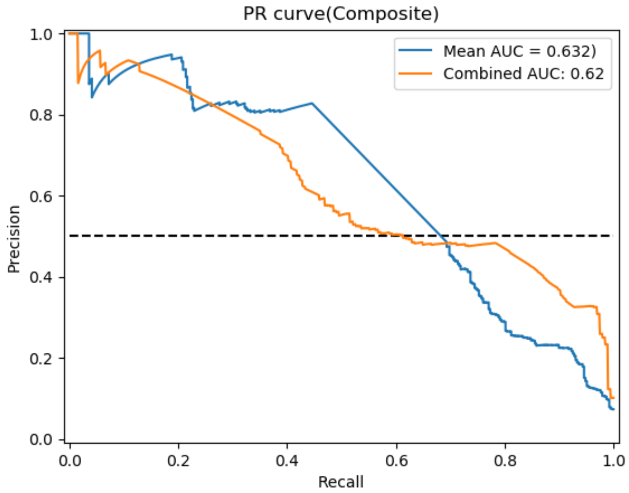

PR curve(Medium)
Mean AUC = 0.424)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


PR curve(Strain)
Mean AUC = 0.712)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

The ROC plots include the micro-average of each k-fold and the mean of these curves as the overall ROC curve. The AUC for strain is 0.91, AUC for medium is 0.84, AUC for stress is 0.81, and the AUC for gene perturbation is 0.97. The AUPRC for strain is 0.712, AUPRC for medium is 0.424, AUPRC for stress is 0.473, and the AUPRC for gene perturbation is 0.823. Looking at the ROC plots, we can see that some folds of the cross validation generate bad results. This has to do with the class imbalance in the dataset and due to the difference in frequency, there was no good way to include each label in each fold which resulted in poor performance for one or two folds.

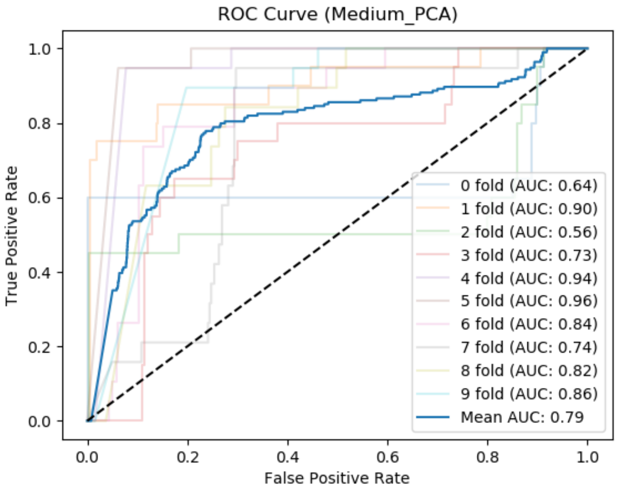
5. I used a similar setup to Q4 except I had to create a list of tuples containing medium and stress and used a Multi Label Binarizer instead.

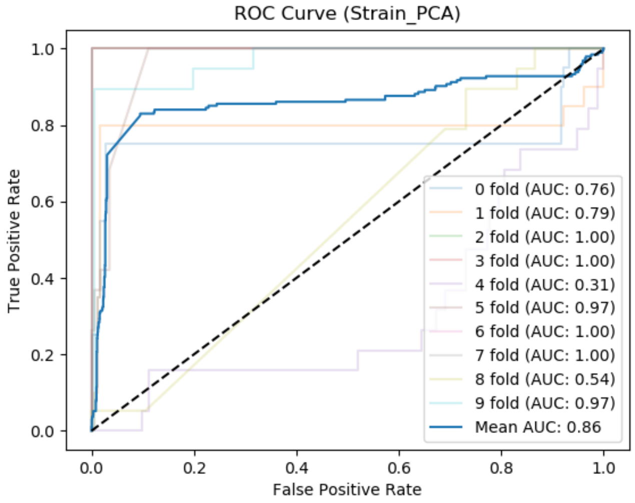


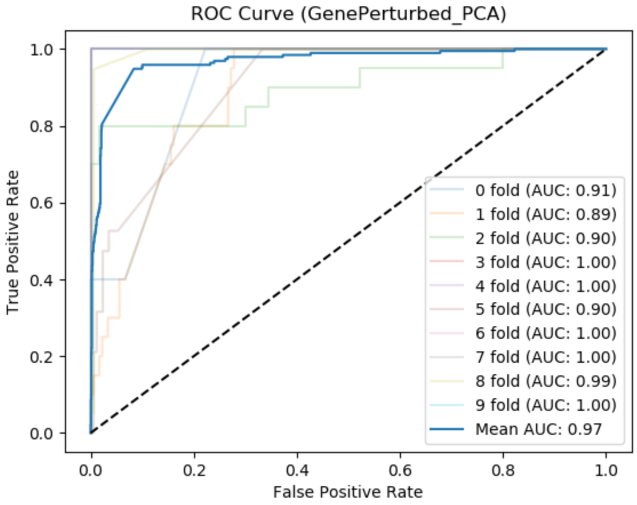


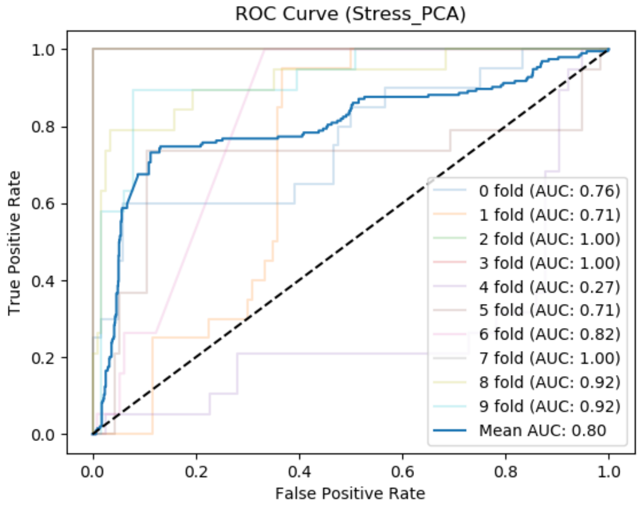
The AUC is 0.91 and the AUPRC is 0.632. The baseline prediction performance is the combination of the medium and environmental perturbation ROC and PR curves. Based on the ROC curves it is not exactly clear but the composite model appears to be slightly better even though the AUC is similar. The PR curves show that the composite AUC is higher and generally performs better than the baseline.

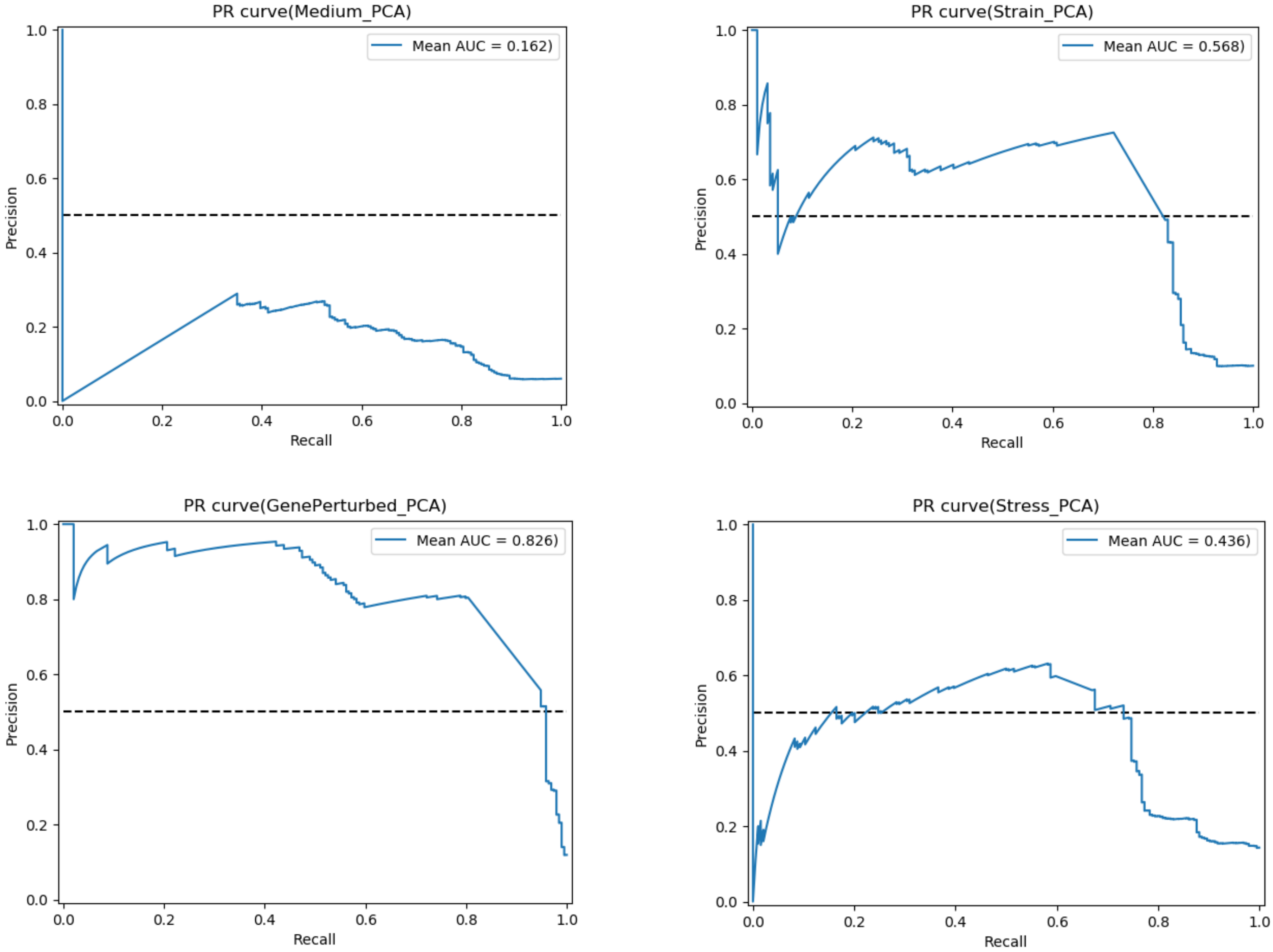
6. I performed PCA on the dataset and fed it into my function from Q4. The AUC for strain is 0.86, AUC for medium is 0.79, AUC for stress is 0.80, AUC for gene perturbation is 0.97. The AUPRC for strain is 0.568, AUPRC for medium is 0.162, AUPRC for stress is 0.436, and the AUPRC for gene perturbation is 0.826. The PCs do retain most of the classification performance while significantly reducing dimensionality. The ROC curves also show the class imbalance within the dataset where the model performance is subpar with 3 of the graphs containing at least one fold heavily underperforming.










PR curve(Medium_PCA)
Mean AUC = 0.162)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR curve(Strain_PCA)
Mean AUC = 0.568)
Precision
Recall
PR curve(GenePerturbed_PCA)
Mean AUC = 0.826)
Precision
Recall
PR curve(Stress_PCA)
Mean AUC = 0.436)
Precision
Recall