Goal:

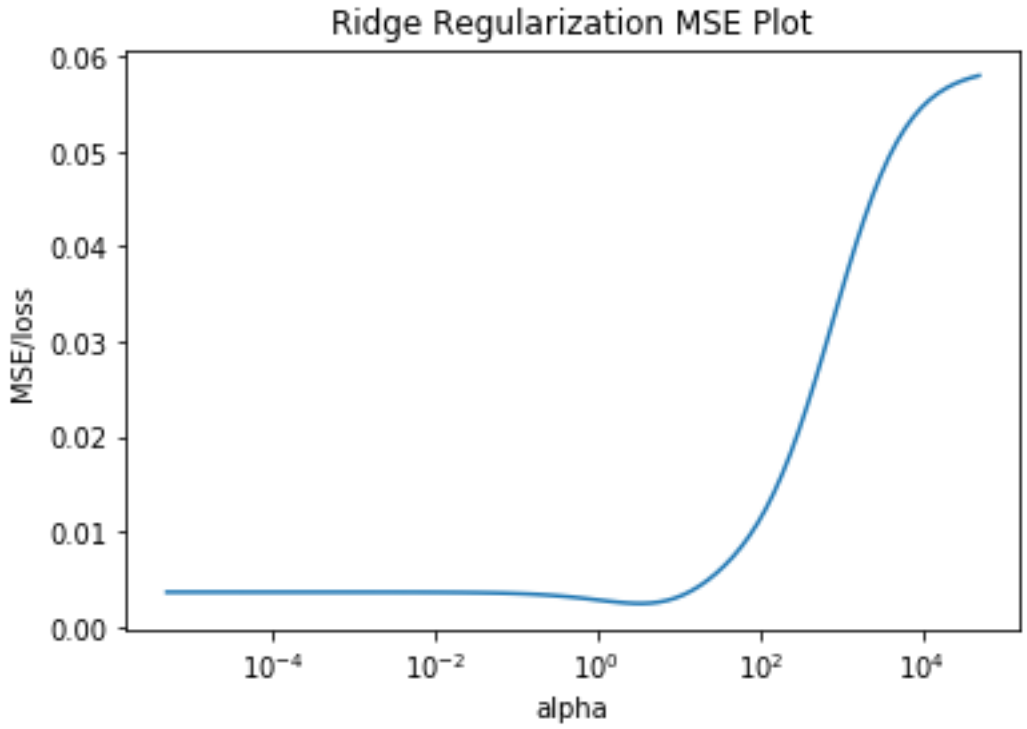
- Use regularized regression to select the informative features
- Calculate the confidence interval of the prediction using the bootstrapping method
- Create Support Vector Machine Classifier and evaluate its performance using ROC (Receiver Operating Characteristic) and PR (Precision – Recall) Curve
- Perform Principal Component Analysis and evaluate its performance using ROC and PR Curve

Problem 1

Create a predictor of the bacterial growth attribute by using only the expression of the genes as attributes. Use regularized regression technique to select genes that are informative for this task. Report the number of features that have non-zero coefficients and the 10-fold cross-validation generalization error of the technique.

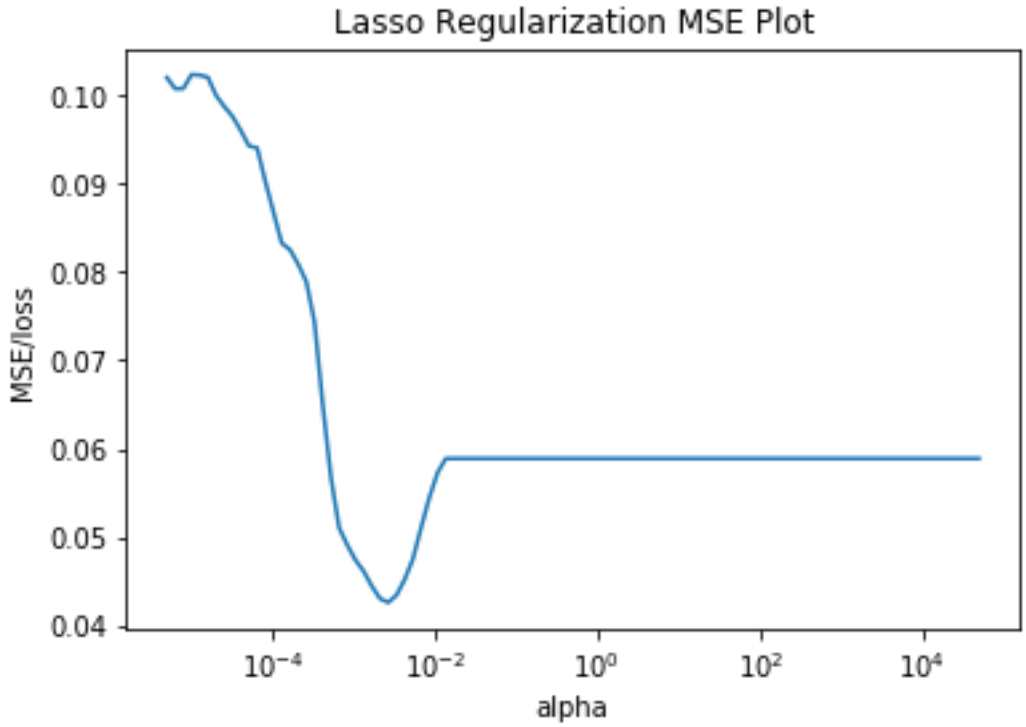
The goal of regularization is to prevent the model from overfitting the training sample. It uses the tuning parameter(lambda) to change the model complexity.

For ridge regularization, there is a slight decrease in the error rate after we increase alpha. However, the error starts to increase dramatically after it reaches its minimum. Using 10-fold cross validation, I found that the best alpha is 9.15 and its MSE (generalization error) is 0.003049. The number of features that have non-zero coefficients are 4433.



For lasso regularization, there is a significant decrease in MSE as I increase alpha. However, the MSE starts to increase back after it reaches its minimum and remains consistent after that. Using

10-fold cross validation, I found that the best alpha is 0.00017 and its MSE (generalization error) is 0.0133. The number of features that have non-zero coefficients are 98.



After comparing the results return from the lasso and ridge regularization, I have decided to use Lasso because it reduces much more features, which is more suitable in this case because our dataset is relatively small.

Problem 2

Extend the predictor and calculate the confidence interval of the prediction by using the bootstrapping method.

Bootstrapping is a method that resample the data with replacement.

Assumptions:
1) Each sample is randomly selected from the population and the selection is independent and unbiased.
2) The sample distribution is a good approximation to the population distribution and the population is infinite.

Problem 3

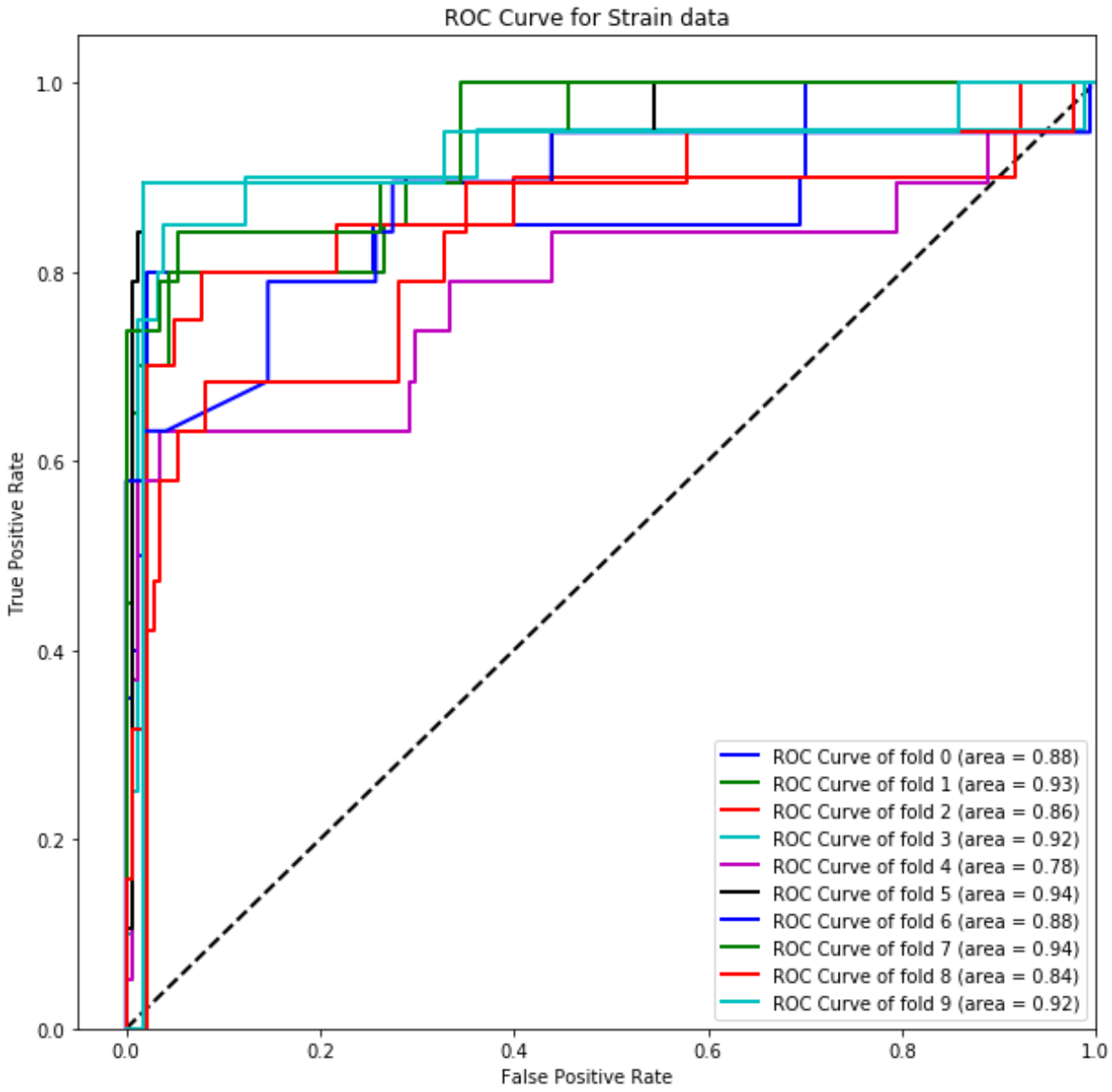
The 95% confidence interval of the predicted growth for a bacterium whose genes are expressed exactly at the mean expression value by using the bootstrapping method is 0.3942 and 0.4351.

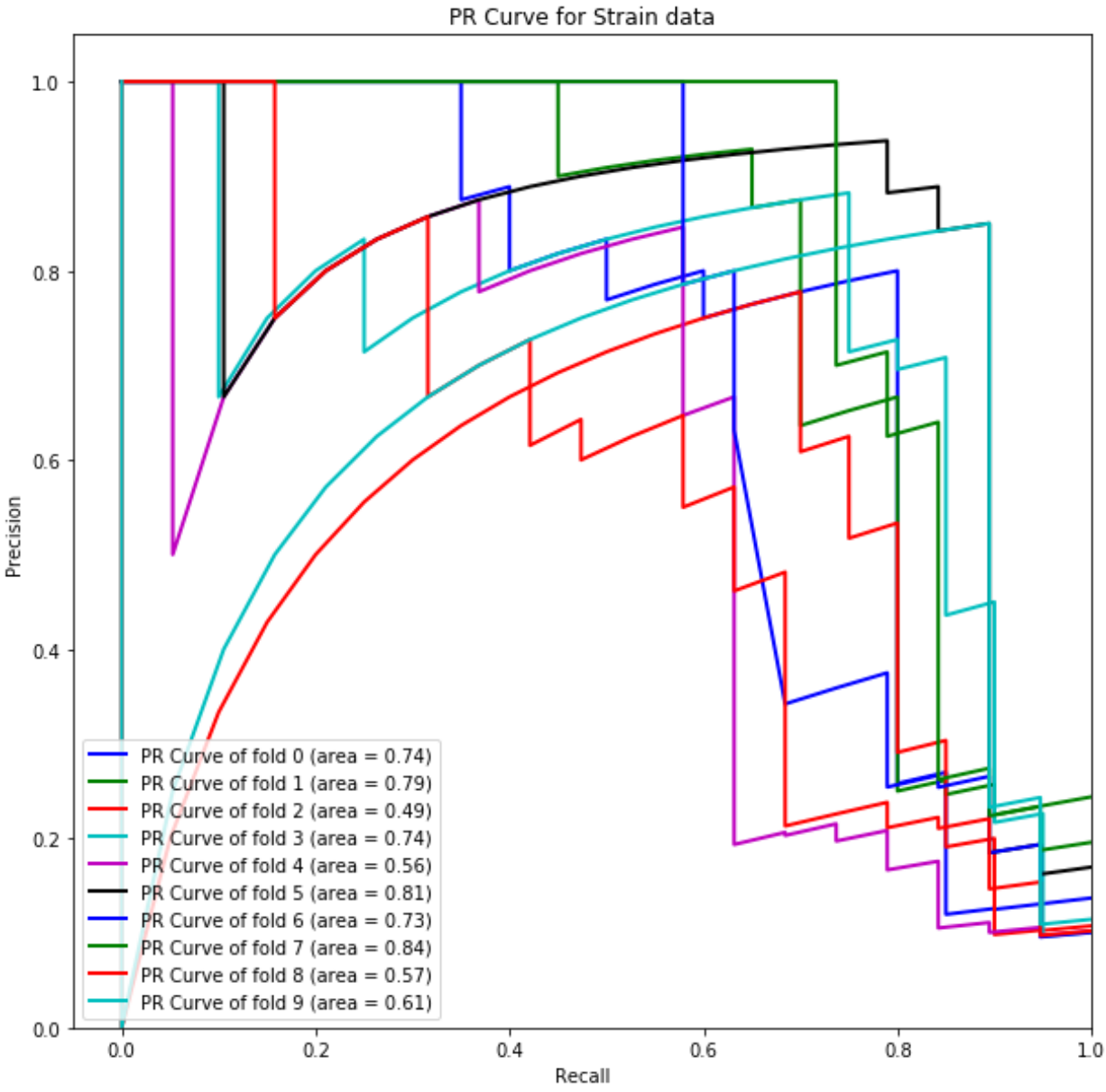
Problem 4

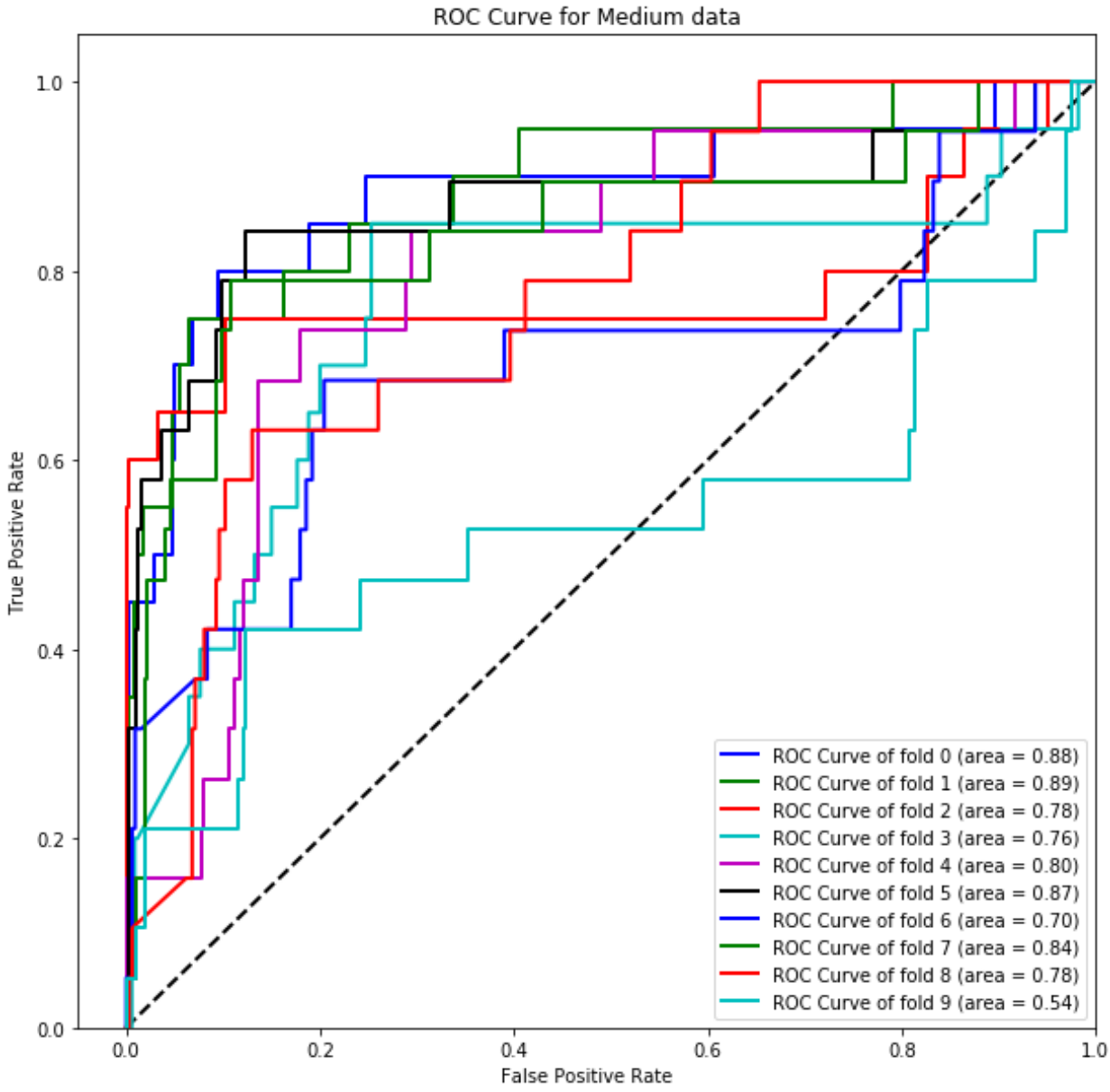
Create four separate SVM classifiers to categorize the strain type, medium type, environmental and gene perturbation, given all the gene transcriptional profiles.

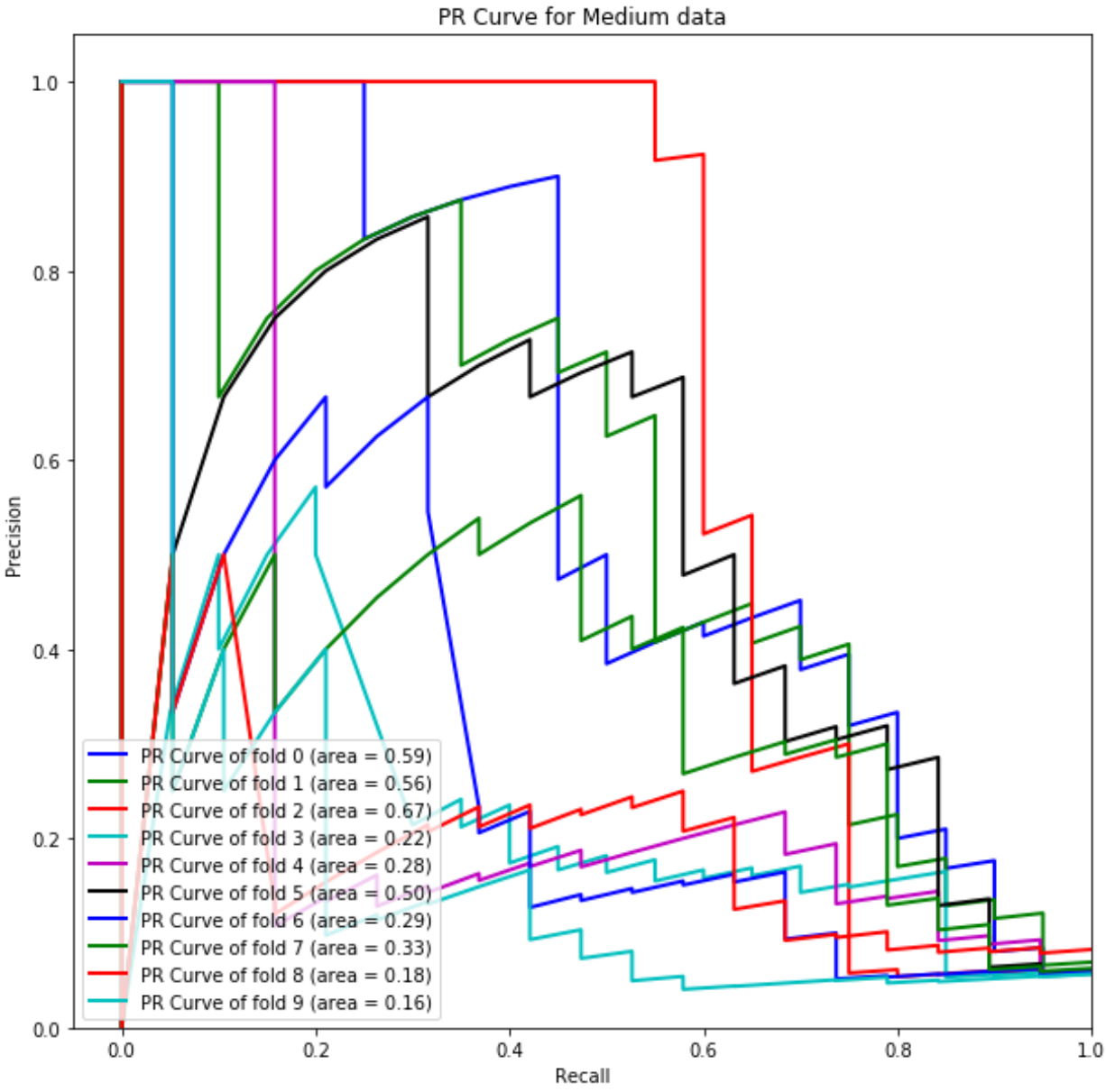
The number of features for all 4 classifiers are 98 because I used the non-zero weighted features from the lasso regularization as X.

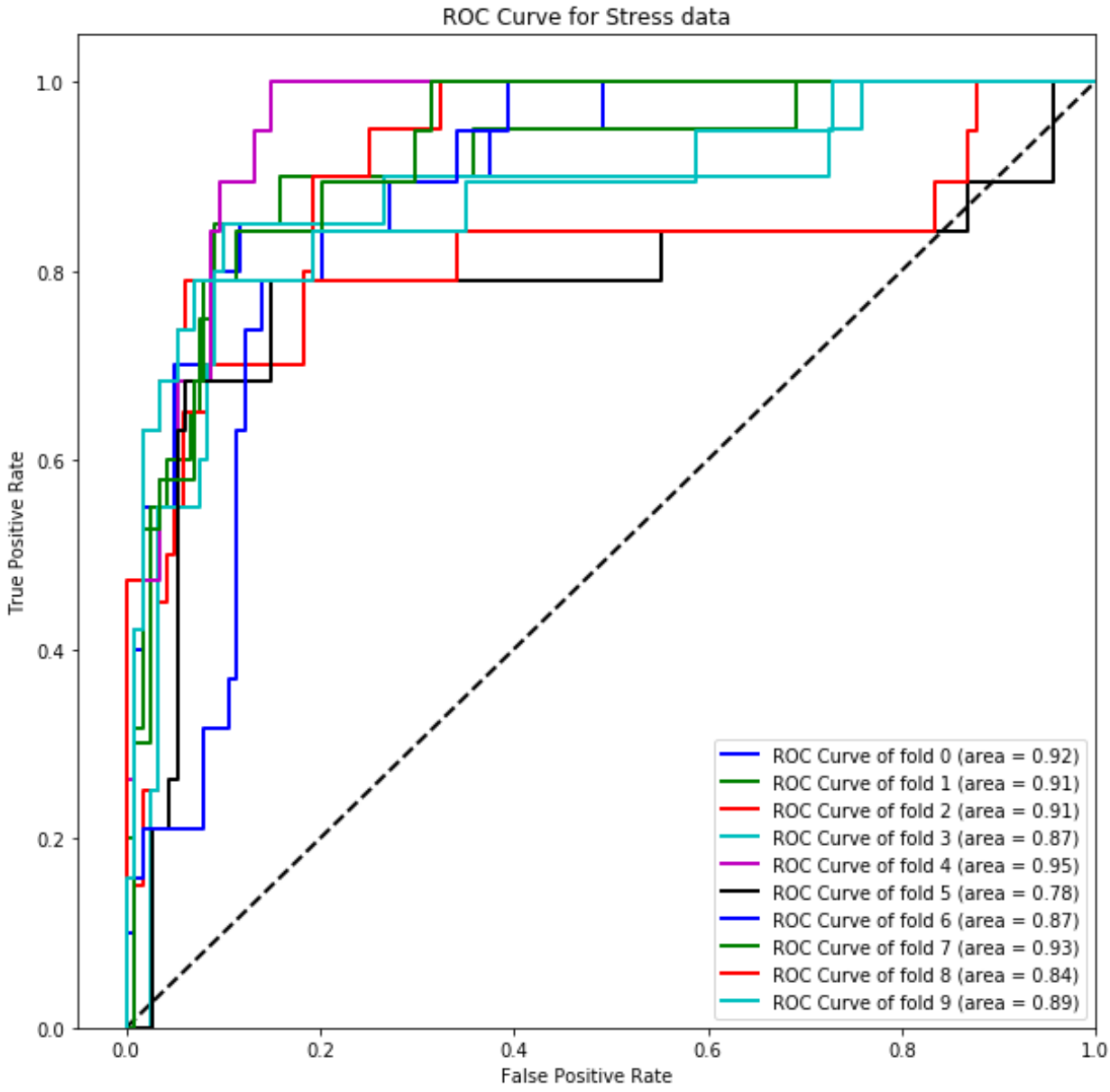
The classifiers do a pretty good job in predicting GenePerturbed, followed by Stress and Strain. However, it does not perform too well for the Medium feature.

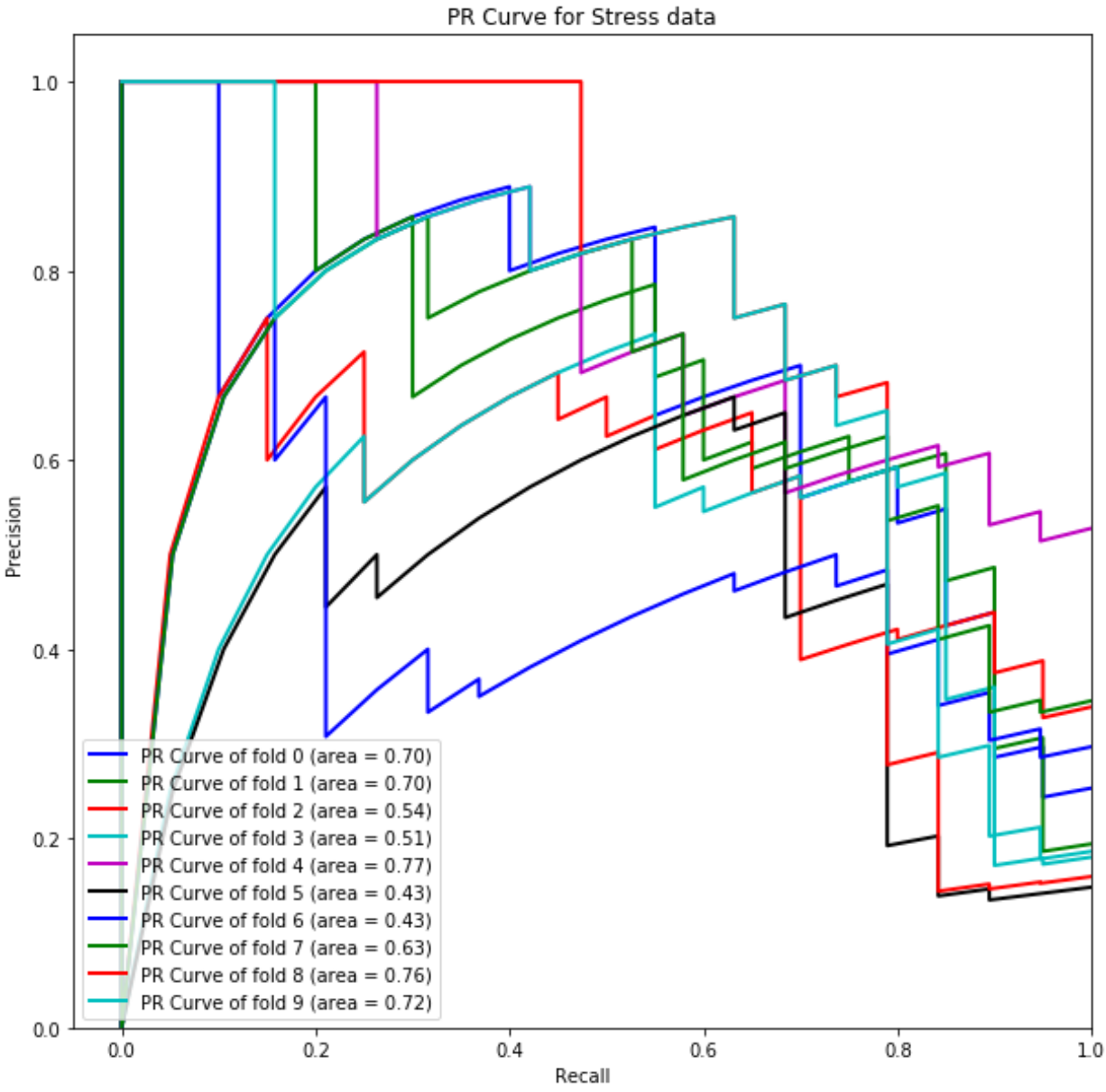


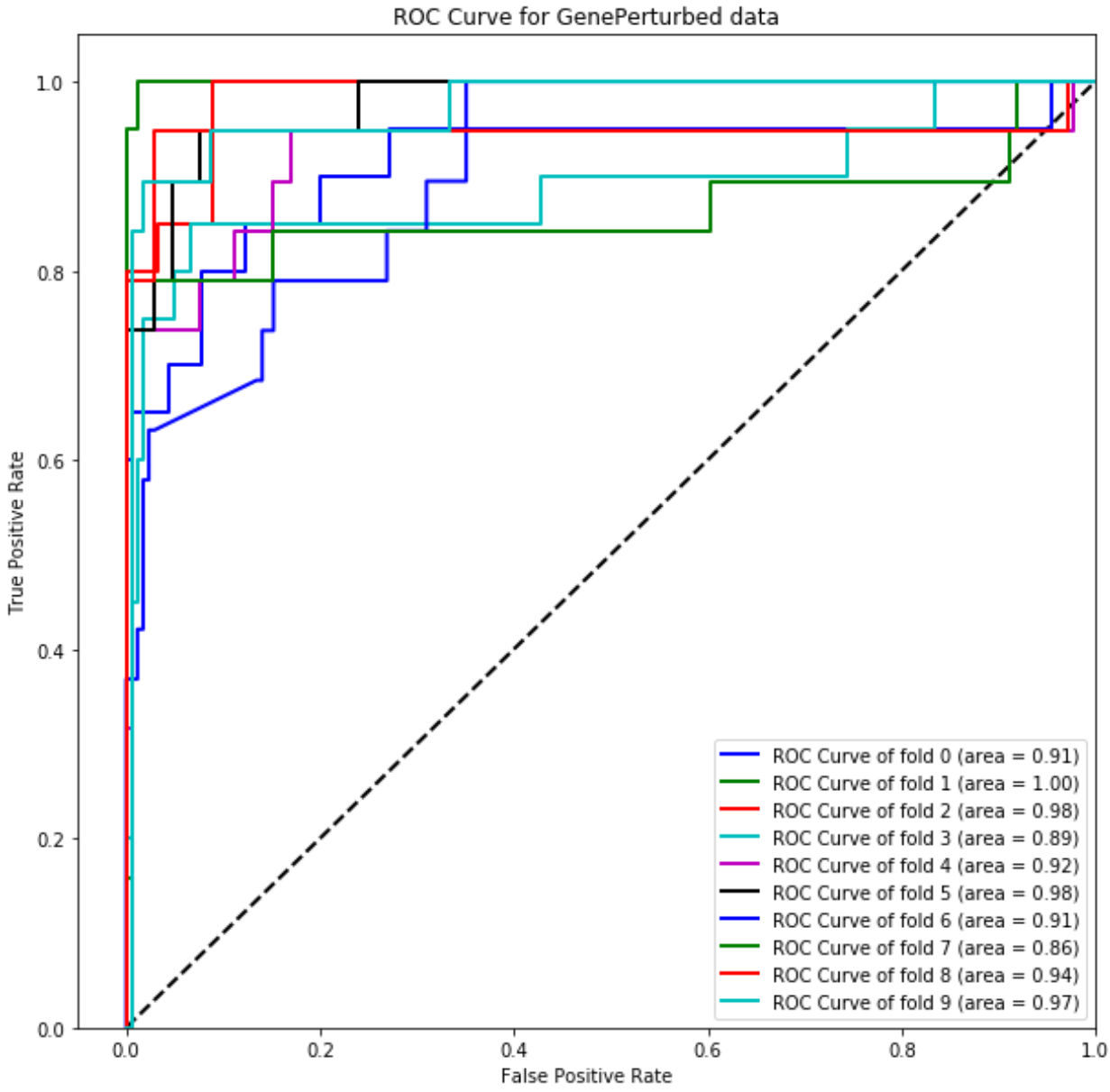

PR Curve for Strain data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.74)
PR Curve of fold 1 (area = 0.79)
PR Curve of fold 2 (area = 0.49)
PR Curve of fold 3 (area = 0.74)
PR Curve of fold 4 (area = 0.56)
PR Curve of fold 5 (area = 0.81)
PR Curve of fold 6 (area = 0.73)
PR Curve of fold 7 (area = 0.84)
PR Curve of fold 8 (area = 0.57)
PR Curve of fold 9 (area = 0.61)

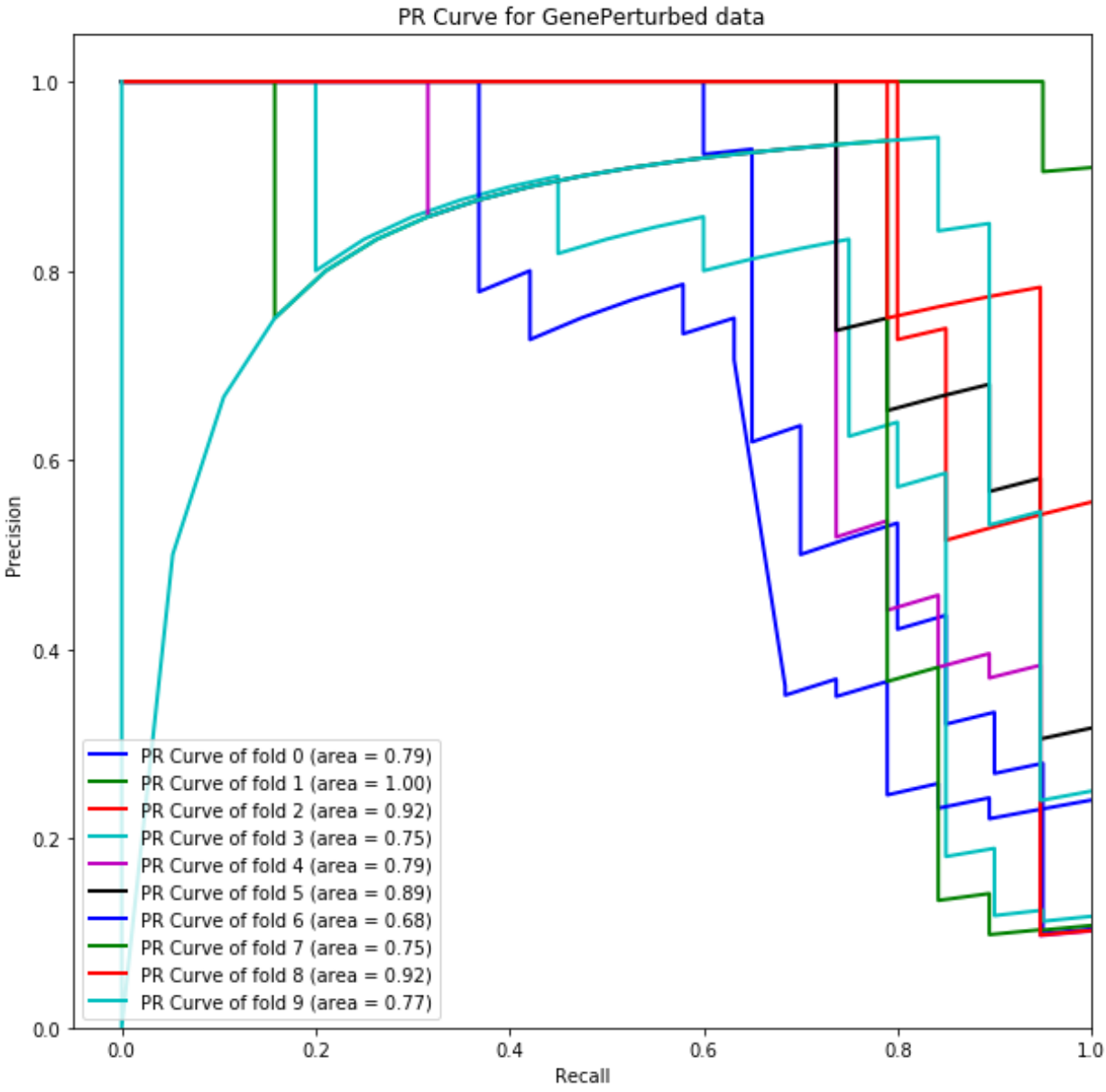

ROC Curve for Medium data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.88)
ROC Curve of fold 1 (area = 0.89)
ROC Curve of fold 2 (area = 0.78)
ROC Curve of fold 3 (area = 0.76)
ROC Curve of fold 4 (area = 0.80)
ROC Curve of fold 5 (area = 0.87)
ROC Curve of fold 6 (area = 0.70)
ROC Curve of fold 7 (area = 0.84)
ROC Curve of fold 8 (area = 0.78)
ROC Curve of fold 9 (area = 0.54)


PR Curve for Medium data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.59)
PR Curve of fold 1 (area = 0.56)
PR Curve of fold 2 (area = 0.67)
PR Curve of fold 3 (area = 0.22)
PR Curve of fold 4 (area = 0.28)
PR Curve of fold 5 (area = 0.50)
PR Curve of fold 6 (area = 0.29)
PR Curve of fold 7 (area = 0.33)
PR Curve of fold 8 (area = 0.18)
PR Curve of fold 9 (area = 0.16)


ROC Curve for Stress data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.92)
ROC Curve of fold 1 (area = 0.91)
ROC Curve of fold 2 (area = 0.91)
ROC Curve of fold 3 (area = 0.87)
ROC Curve of fold 4 (area = 0.95)
ROC Curve of fold 5 (area = 0.78)
ROC Curve of fold 6 (area = 0.87)
ROC Curve of fold 7 (area = 0.93)
ROC Curve of fold 8 (area = 0.84)
ROC Curve of fold 9 (area = 0.89)


PR Curve for Stress data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.70)
PR Curve of fold 1 (area = 0.70)
PR Curve of fold 2 (area = 0.54)
PR Curve of fold 3 (area = 0.51)
PR Curve of fold 4 (area = 0.77)
PR Curve of fold 5 (area = 0.43)
PR Curve of fold 6 (area = 0.43)
PR Curve of fold 7 (area = 0.63)
PR Curve of fold 8 (area = 0.76)
PR Curve of fold 9 (area = 0.72)

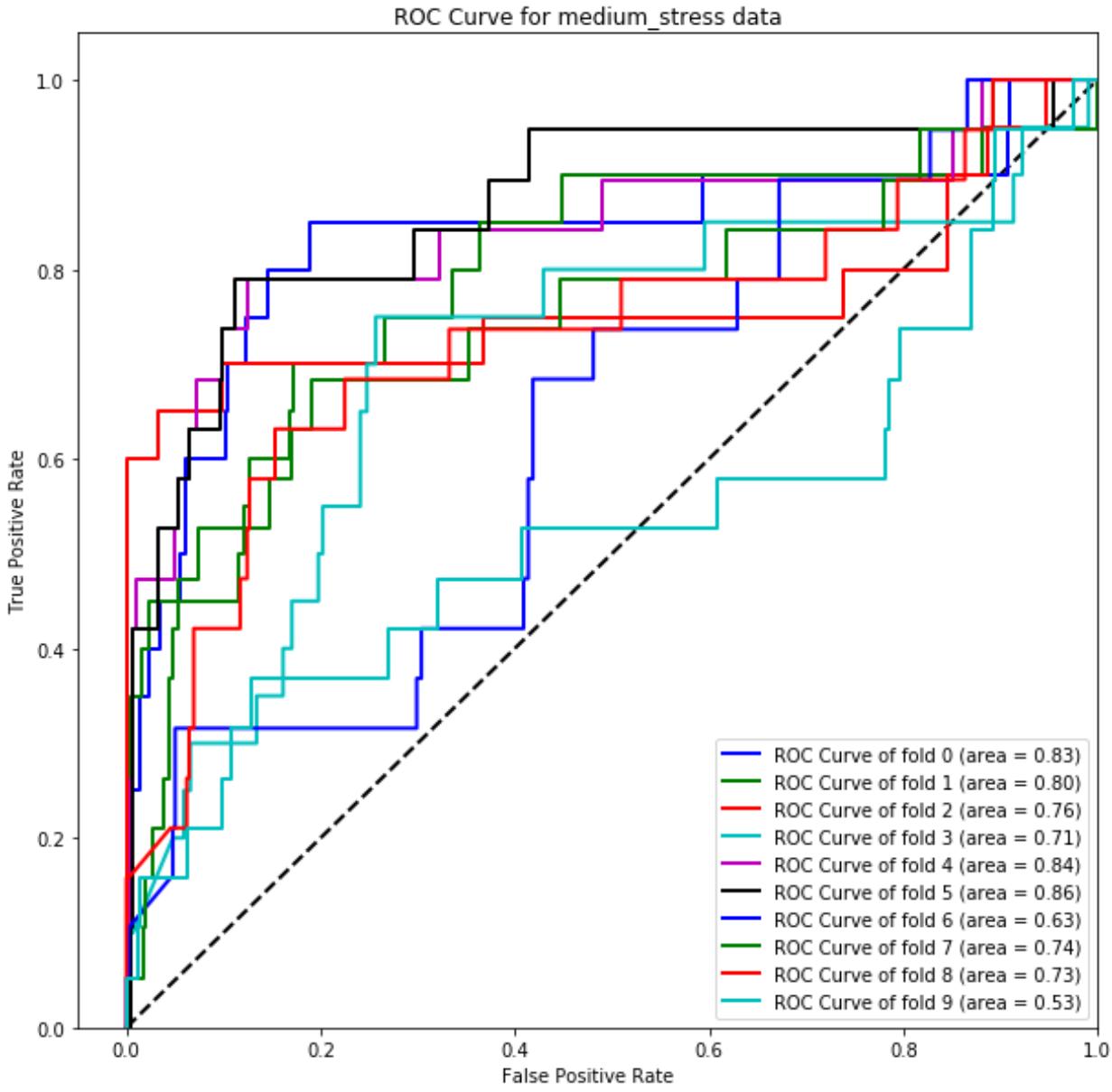

ROC Curve for GenePerturbed data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.91)
ROC Curve of fold 1 (area = 1.00)
ROC Curve of fold 2 (area = 0.98)
ROC Curve of fold 3 (area = 0.89)
ROC Curve of fold 4 (area = 0.92)
ROC Curve of fold 5 (area = 0.98)
ROC Curve of fold 6 (area = 0.91)
ROC Curve of fold 7 (area = 0.86)
ROC Curve of fold 8 (area = 0.94)
ROC Curve of fold 9 (area = 0.97)

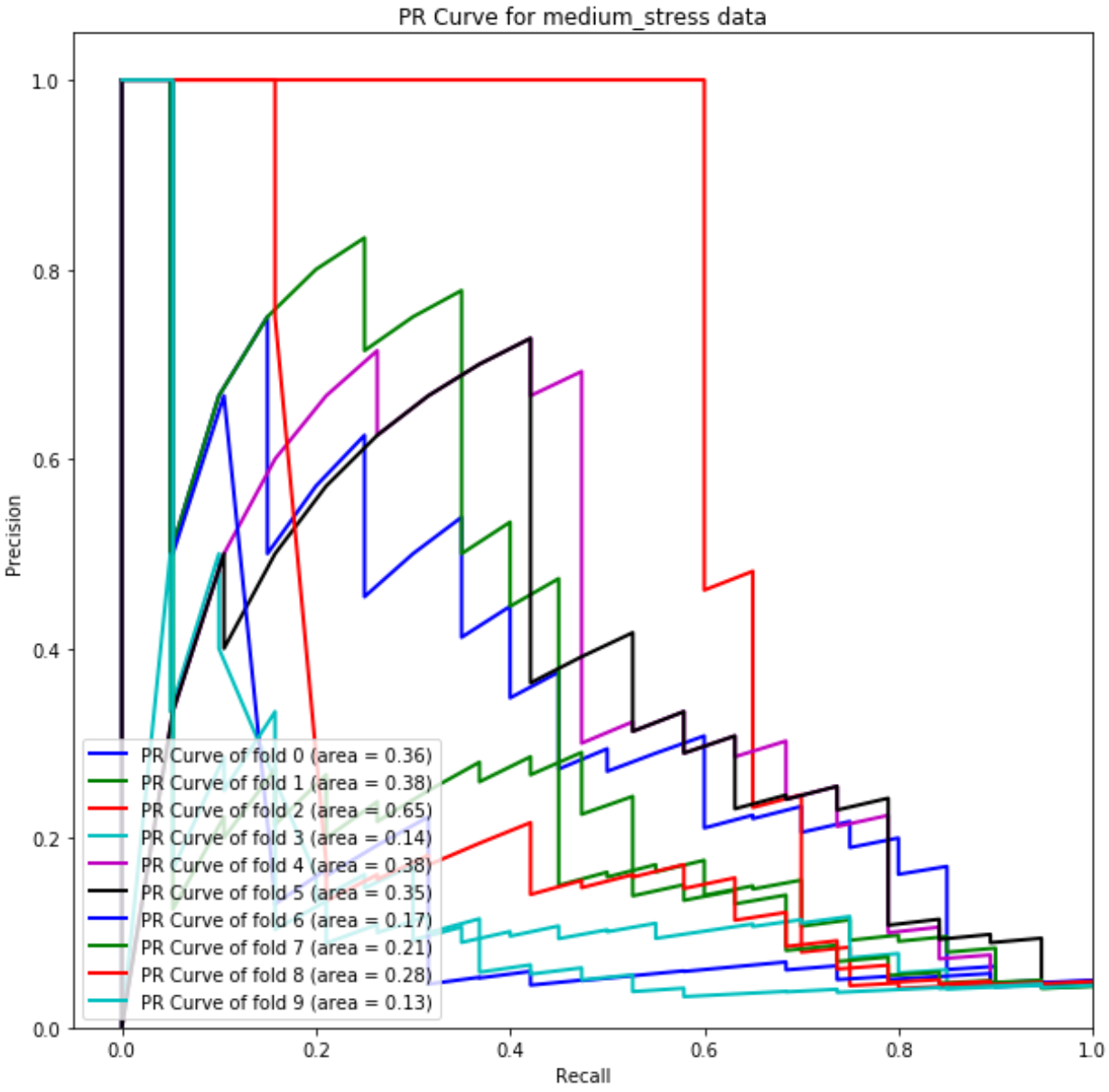


Problem 5

Create one composite SVM classifier to simultaneously predict medium and environmental perturbations and report the 10-fold cross-validation AUC/AUPRC value.

This classifier performs worse than the two individual classifiers together for these predictions. The null hypothesis is that the combined classifier performs better or equal to the two individual classifiers. Our result proves that our null hypothesis is incorrect.

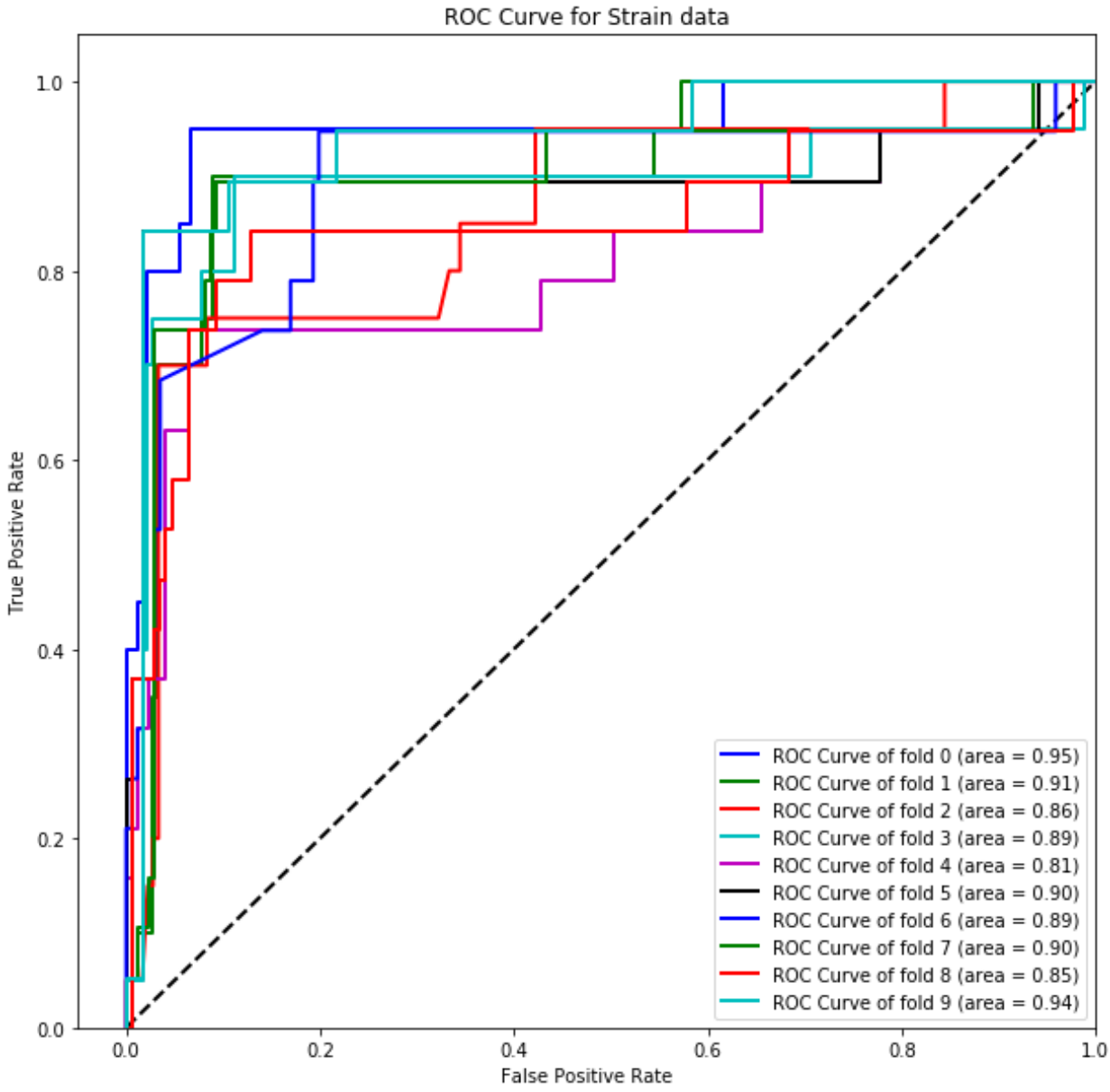

ROC Curve for medium_stress data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.83)
ROC Curve of fold 1 (area = 0.80)
ROC Curve of fold 2 (area = 0.76)
ROC Curve of fold 3 (area = 0.71)
ROC Curve of fold 4 (area = 0.84)
ROC Curve of fold 5 (area = 0.86)
ROC Curve of fold 6 (area = 0.63)
ROC Curve of fold 7 (area = 0.74)
ROC Curve of fold 8 (area = 0.73)
ROC Curve of fold 9 (area = 0.53)

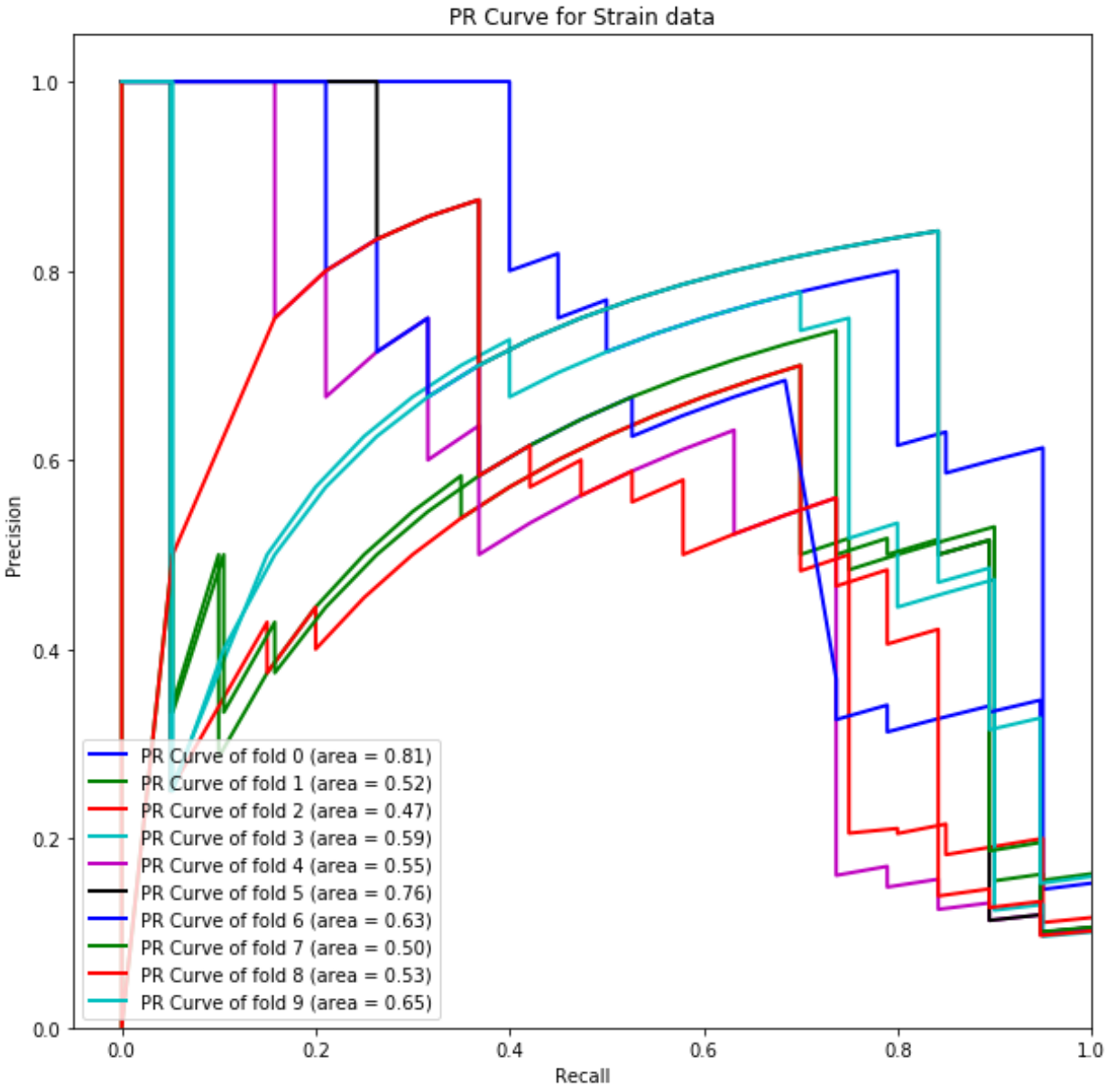


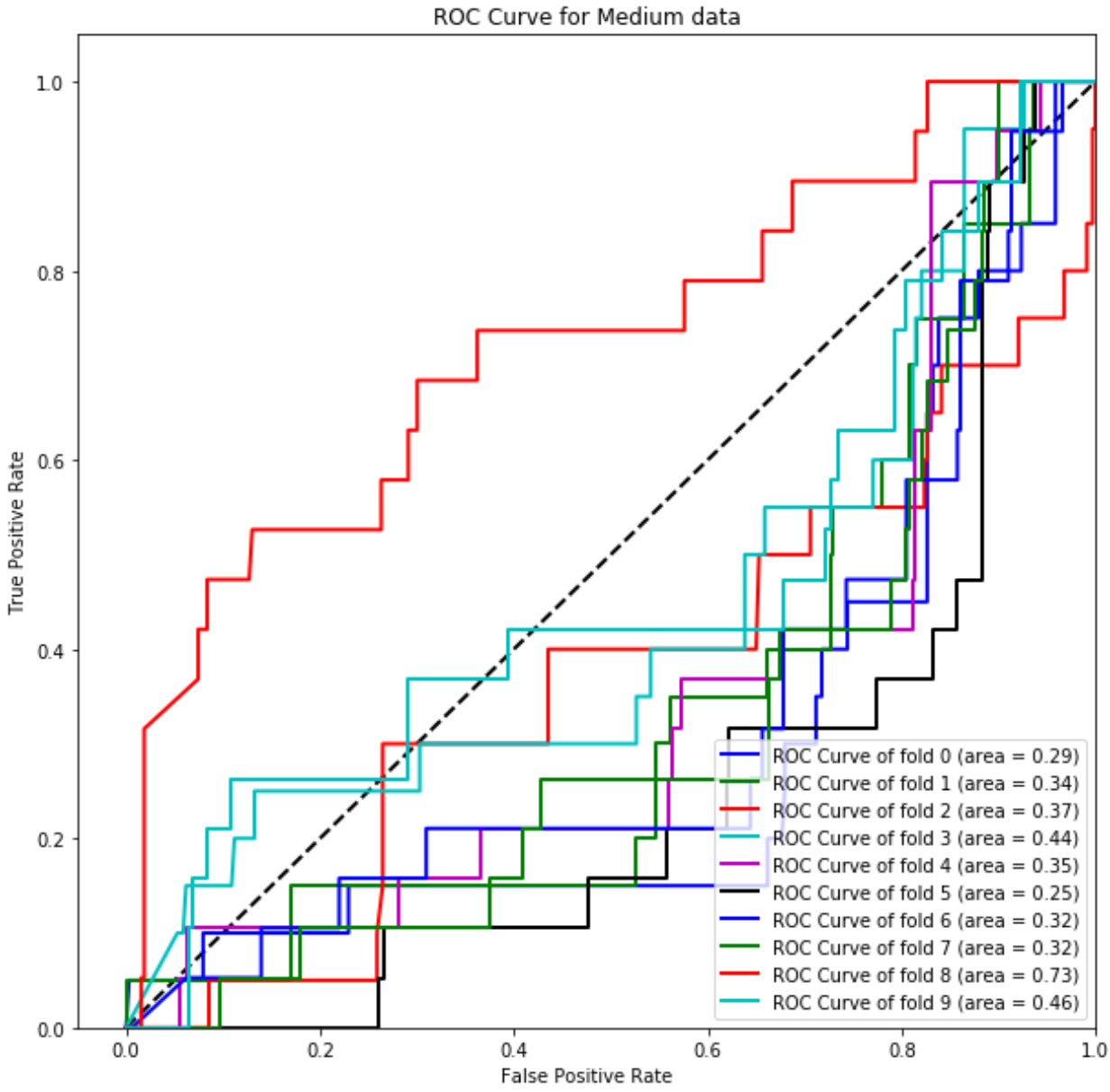
Problem 6

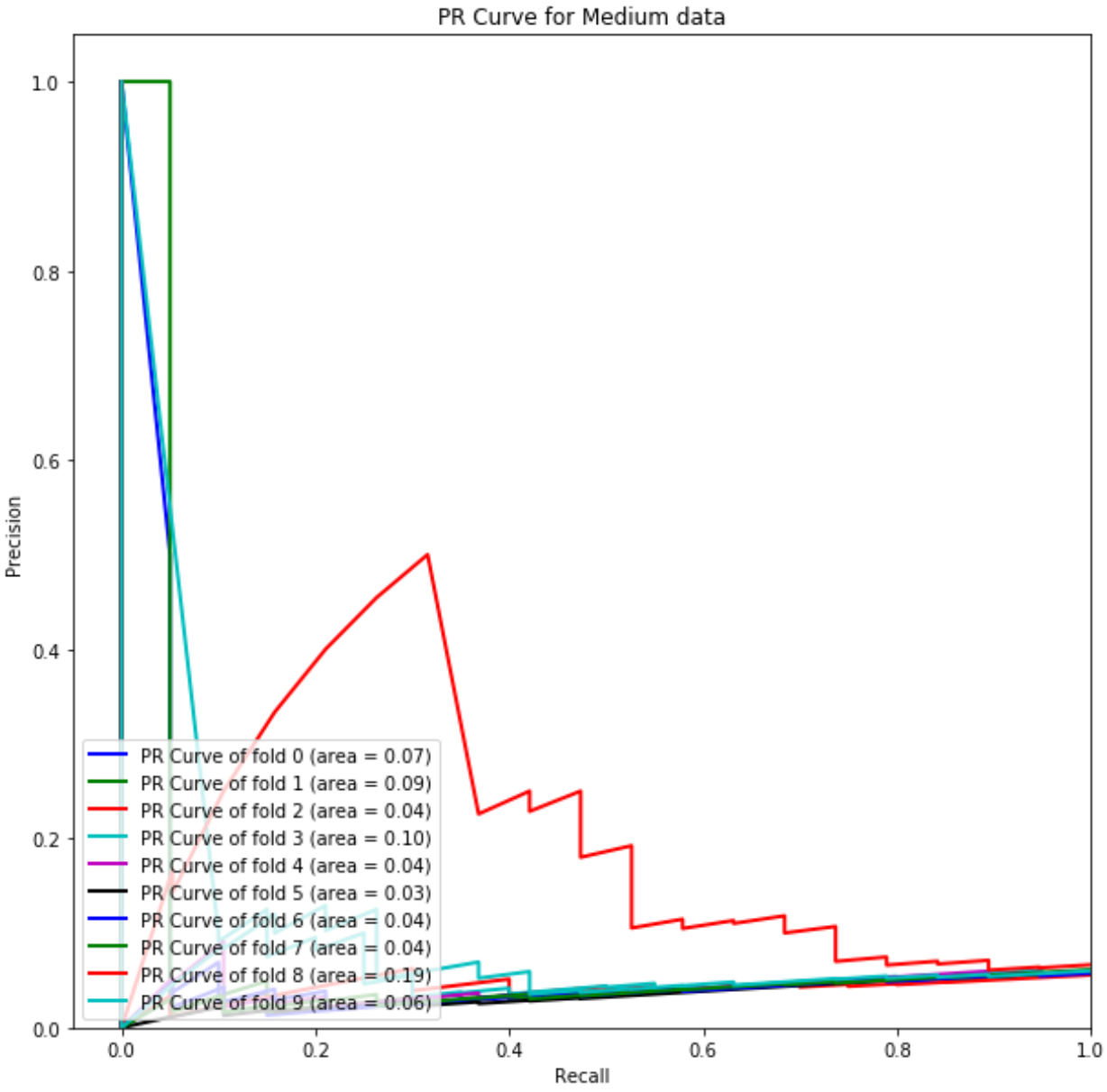
Perform Principal Component Analysis, keeping only the 3 Principal Components (PCs) as features for the SVM classifier.

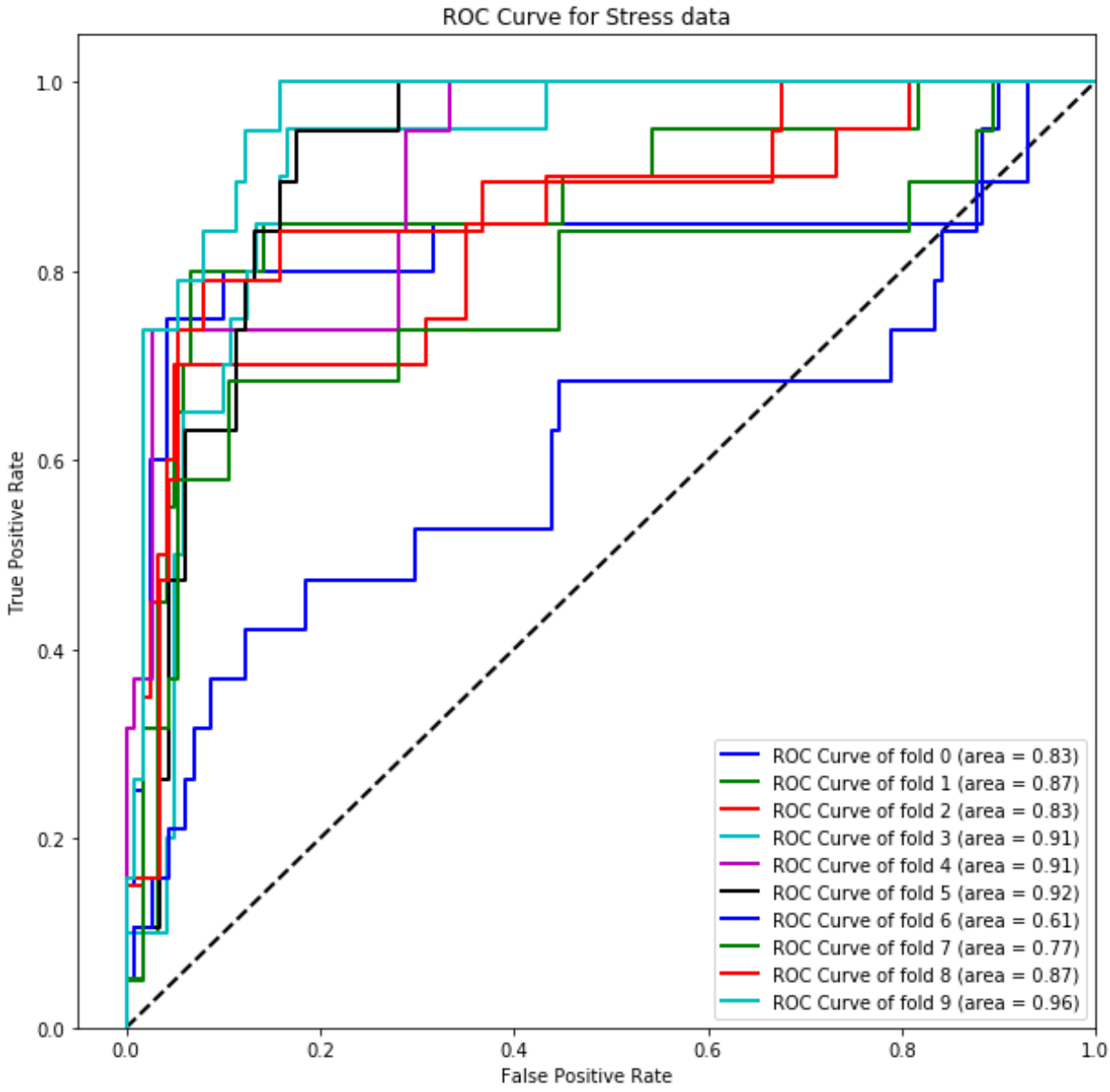
The PCs retain most of the classification performance while reducing the dimensionality because AUC/AUPRC value for Strain, Stress and GenePerturbed are still relatively high given we only have three principal components. However, AUC/AUPRC value for Medium drops significantly.

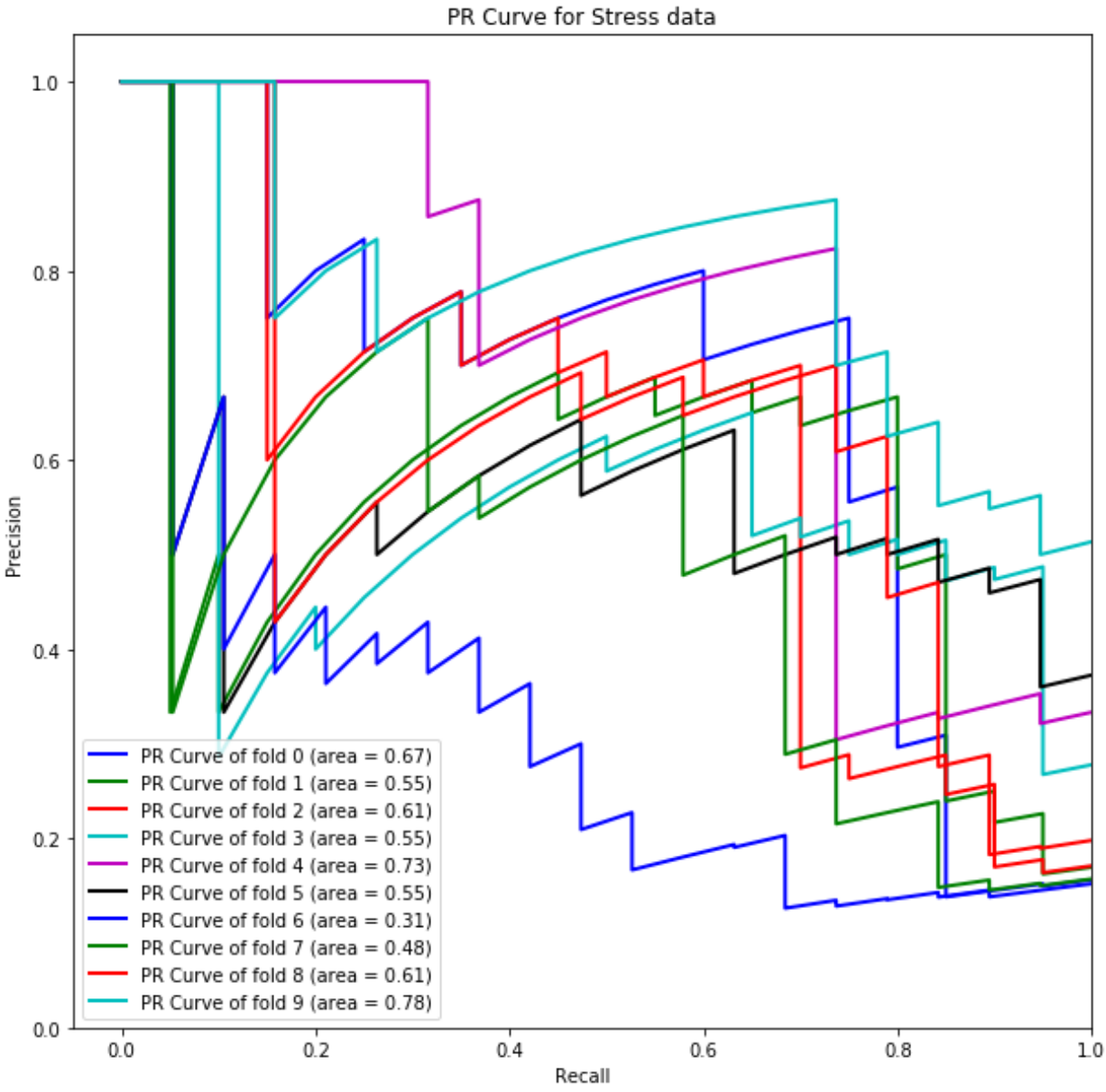

ROC Curve for Strain data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.95)
ROC Curve of fold 1 (area = 0.91)
ROC Curve of fold 2 (area = 0.86)
ROC Curve of fold 3 (area = 0.89)
ROC Curve of fold 4 (area = 0.81)
ROC Curve of fold 5 (area = 0.90)
ROC Curve of fold 6 (area = 0.89)
ROC Curve of fold 7 (area = 0.90)
ROC Curve of fold 8 (area = 0.85)
ROC Curve of fold 9 (area = 0.94)

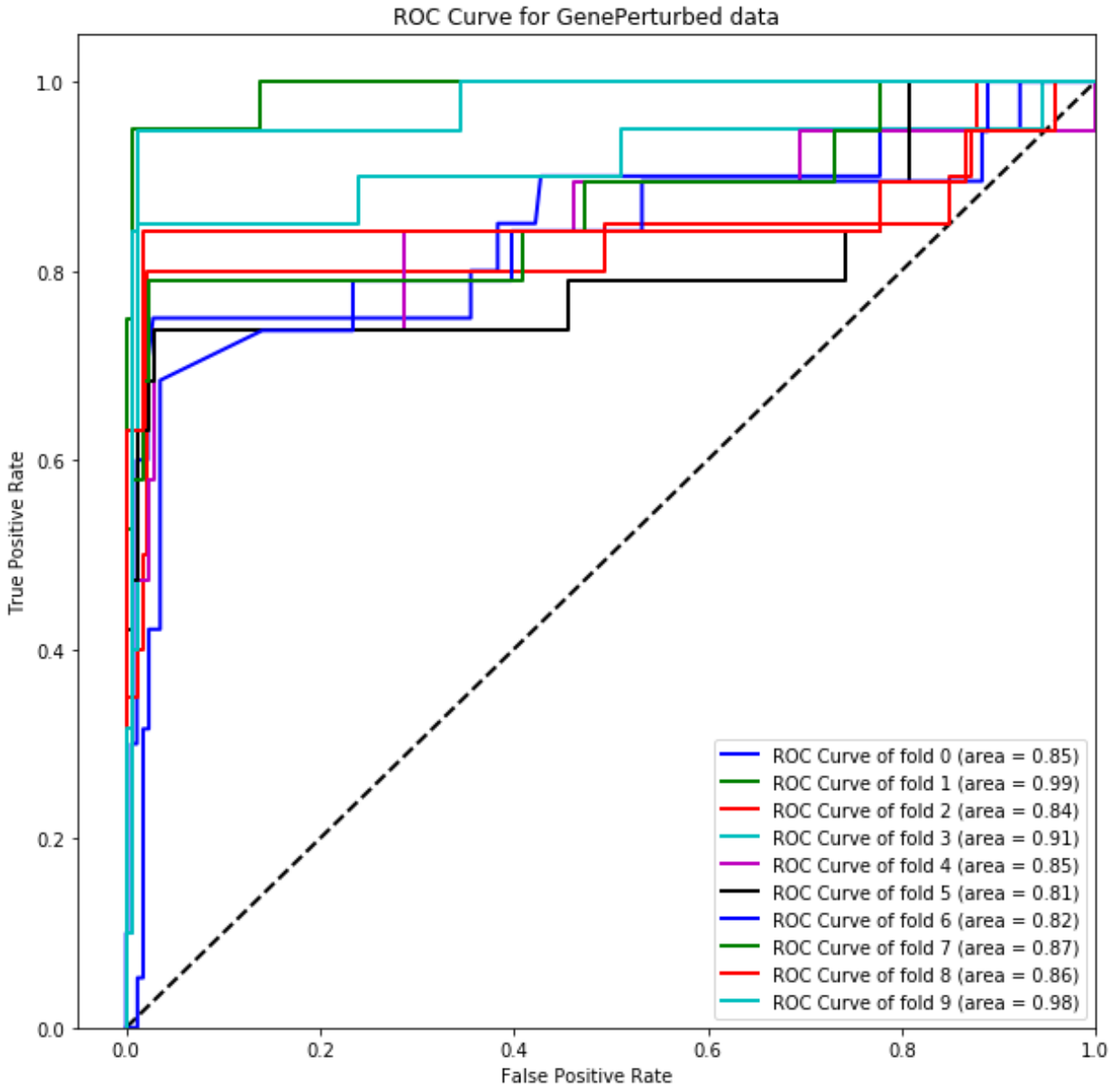

PR Curve for Strain data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.81)
PR Curve of fold 1 (area = 0.52)
PR Curve of fold 2 (area = 0.47)
PR Curve of fold 3 (area = 0.59)
PR Curve of fold 4 (area = 0.55)
PR Curve of fold 5 (area = 0.76)
PR Curve of fold 6 (area = 0.63)
PR Curve of fold 7 (area = 0.50)
PR Curve of fold 8 (area = 0.53)
PR Curve of fold 9 (area = 0.65)

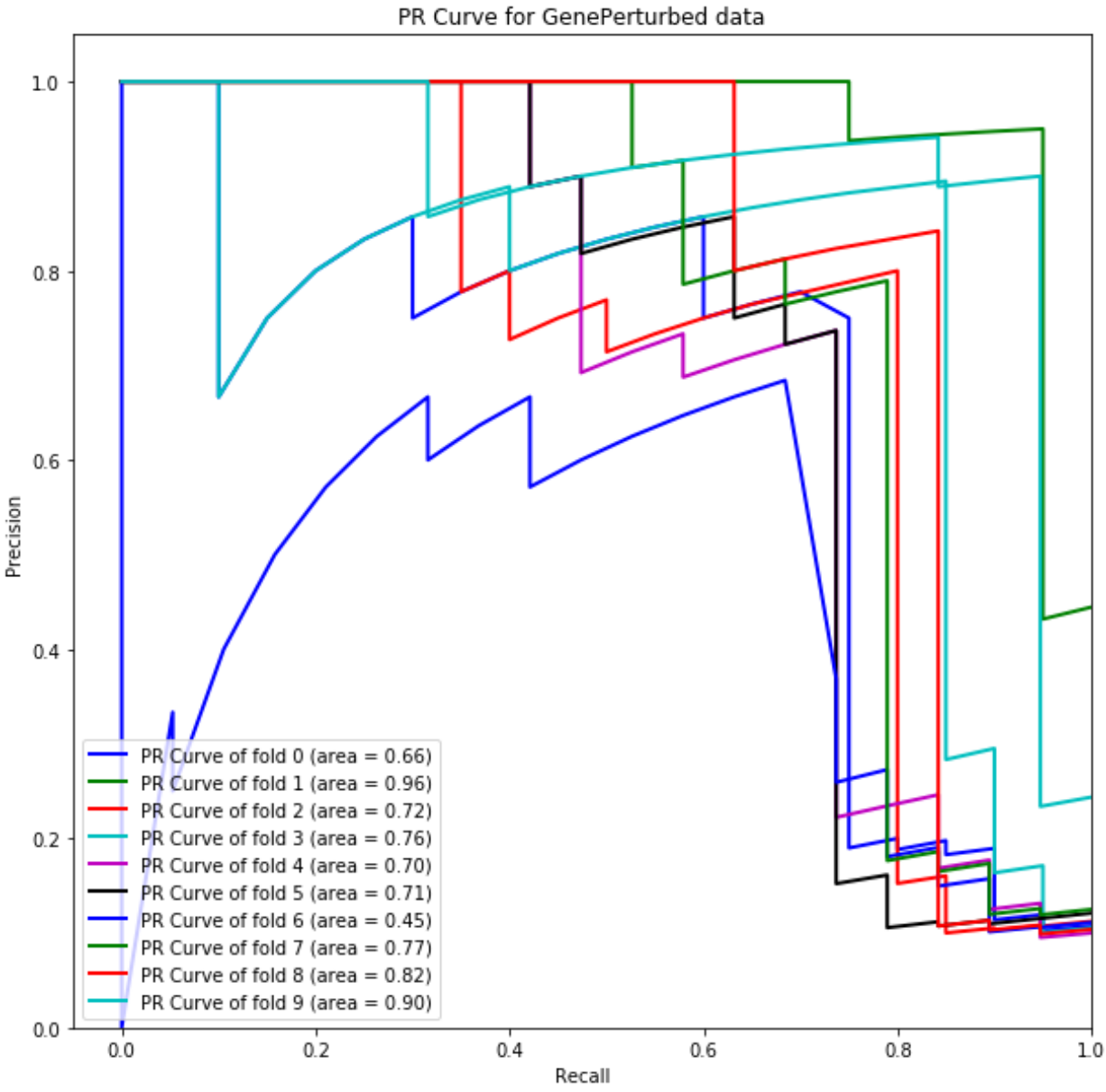

ROC Curve for Medium data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.29)
ROC Curve of fold 1 (area = 0.34)
ROC Curve of fold 2 (area = 0.37)
ROC Curve of fold 3 (area = 0.44)
ROC Curve of fold 4 (area = 0.35)
ROC Curve of fold 5 (area = 0.25)
ROC Curve of fold 6 (area = 0.32)
ROC Curve of fold 7 (area = 0.32)
ROC Curve of fold 8 (area = 0.73)
ROC Curve of fold 9 (area = 0.46)


PR Curve for Medium data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.07)
PR Curve of fold 1 (area = 0.09)
PR Curve of fold 2 (area = 0.04)
PR Curve of fold 3 (area = 0.10)
PR Curve of fold 4 (area = 0.04)
PR Curve of fold 5 (area = 0.03)
PR Curve of fold 6 (area = 0.04)
PR Curve of fold 7 (area = 0.04)
PR Curve of fold 8 (area = 0.19)
PR Curve of fold 9 (area = 0.06)


ROC Curve for Stress data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.83)
ROC Curve of fold 1 (area = 0.87)
ROC Curve of fold 2 (area = 0.83)
ROC Curve of fold 3 (area = 0.91)
ROC Curve of fold 4 (area = 0.91)
ROC Curve of fold 5 (area = 0.92)
ROC Curve of fold 6 (area = 0.61)
ROC Curve of fold 7 (area = 0.77)
ROC Curve of fold 8 (area = 0.87)
ROC Curve of fold 9 (area = 0.96)


PR Curve for Stress data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.67)
PR Curve of fold 1 (area = 0.55)
PR Curve of fold 2 (area = 0.61)
PR Curve of fold 3 (area = 0.55)
PR Curve of fold 4 (area = 0.73)
PR Curve of fold 5 (area = 0.55)
PR Curve of fold 6 (area = 0.31)
PR Curve of fold 7 (area = 0.48)
PR Curve of fold 8 (area = 0.61)
PR Curve of fold 9 (area = 0.78)


ROC Curve for GenePerturbed data
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve of fold 0 (area = 0.85)
ROC Curve of fold 1 (area = 0.99)
ROC Curve of fold 2 (area = 0.84)
ROC Curve of fold 3 (area = 0.91)
ROC Curve of fold 4 (area = 0.85)
ROC Curve of fold 5 (area = 0.81)
ROC Curve of fold 6 (area = 0.82)
ROC Curve of fold 7 (area = 0.87)
ROC Curve of fold 8 (area = 0.86)
ROC Curve of fold 9 (area = 0.98)


PR Curve for GenePerturbed data
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
PR Curve of fold 0 (area = 0.66)
PR Curve of fold 1 (area = 0.96)
PR Curve of fold 2 (area = 0.72)
PR Curve of fold 3 (area = 0.76)
PR Curve of fold 4 (area = 0.70)
PR Curve of fold 5 (area = 0.71)
PR Curve of fold 6 (area = 0.45)
PR Curve of fold 7 (area = 0.77)
PR Curve of fold 8 (area = 0.82)
PR Curve of fold 9 (area = 0.90)