Melody Chang 912110826

ECS171 HW3

1) Refer to part1.py

   For this problem I used Lasso regression. Lasso selects the only some features while reducing the coefficients of others to zero. This property is known as feature selection. I tried out a bunch of values for lambda and found that 0.001 was the most optimal, giving an MSE of 0.037 (which is the generalization error) and number of zero coefficients of 113.

```
Cross Score:  [ 0.0197584   0.01568513  0.02947095  0.12226379  0.00431666  0.03393869
  0.02949795  0.00898991  0.10477187  0.00345881]
MSE: 0.037215215481
Number of non zero coefficients:  113
```

2) Refer to part2.py

   I resampled the dataset with 500 iterations. For each model I can get a prediction Y, and calculate the mean and standard deviation and get the confidence interval for Y. For a 95 confidence interval I got a lower bound of 0.0092 and a upper bound of 0.07919.

```
95.0 confidence interval  0.00923591313045  and  0.0791910058101
```

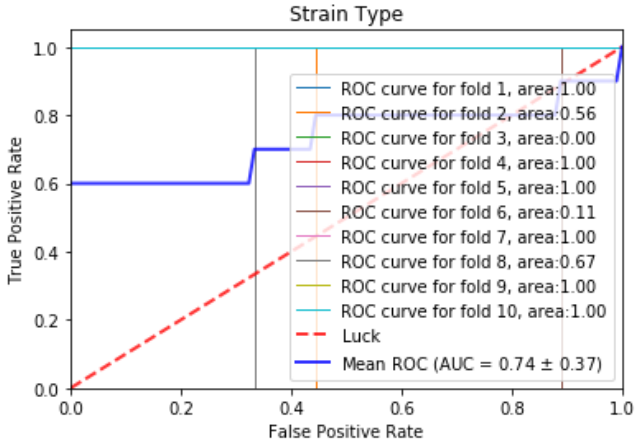
3) Refer to part3.py

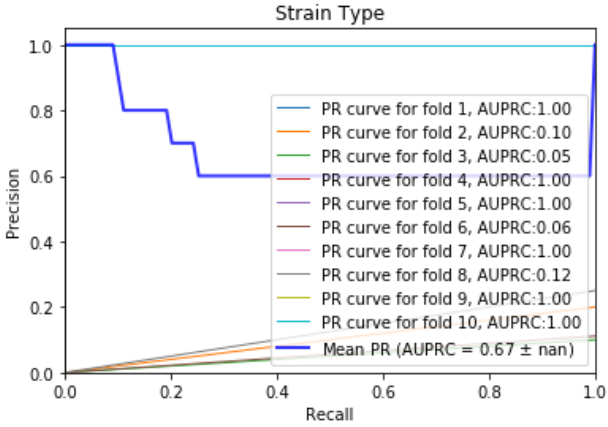
   For this part I found the mean expression value, which is 0.3936.

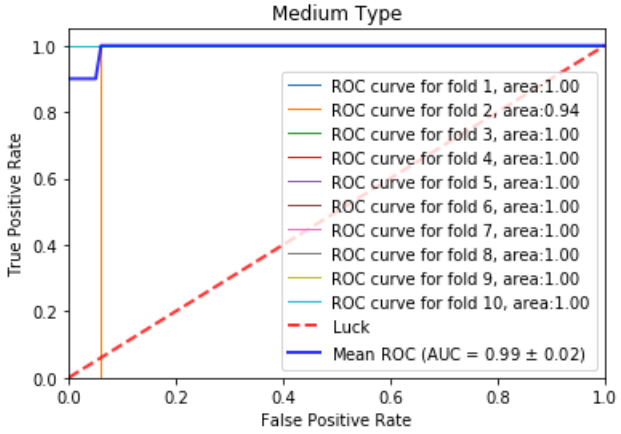
```
Y output: [ 0.39363918]
```

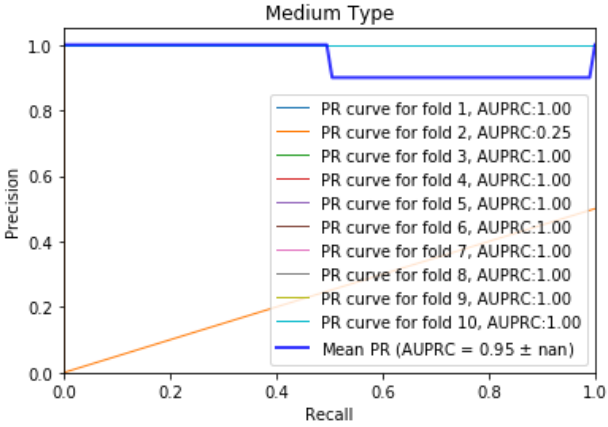
4) Refer to part4_ROC.py and part4_PR.py

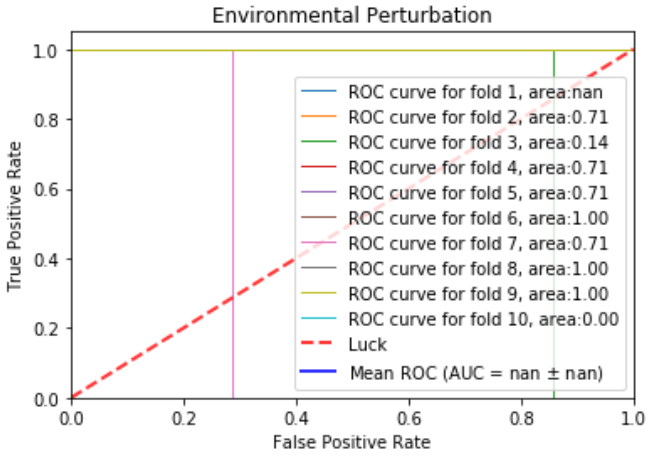
   For this problem I used feature selection with the number of features as 50. The graphs of the 4 classifiers are shown below with the AUC/AUPRC values:

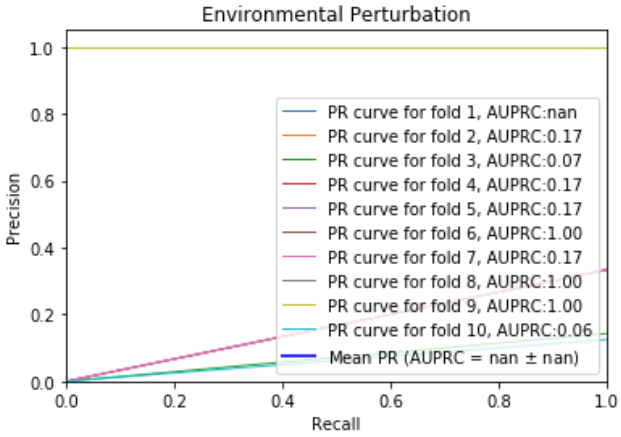

Strain Type
True Positive Rate
False Positive Rate
ROC curve for fold 1, area:1.00
ROC curve for fold 2, area:0.56
ROC curve for fold 3, area:0.00
ROC curve for fold 4, area:1.00
ROC curve for fold 5, area:1.00
ROC curve for fold 6, area:0.11
ROC curve for fold 7, area:1.00
ROC curve for fold 8, area:0.67
ROC curve for fold 9, area:1.00
ROC curve for fold 10, area:1.00
Luck
Mean ROC (AUC = 0.74 ± 0.37)

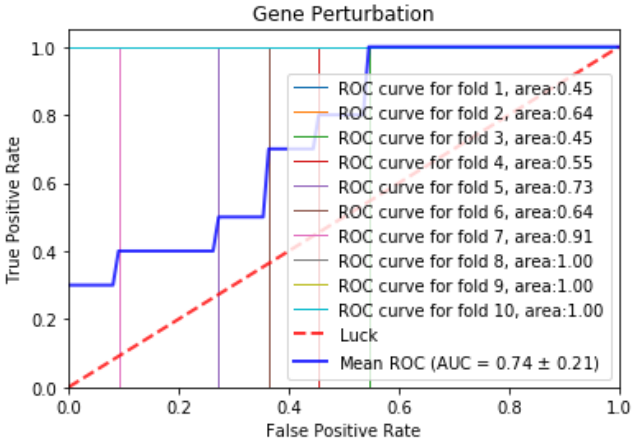

Strain Type
Precision
Recall
PR curve for fold 1, AUPRC:1.00
PR curve for fold 2, AUPRC:0.10
PR curve for fold 3, AUPRC:0.05
PR curve for fold 4, AUPRC:1.00
PR curve for fold 5, AUPRC:1.00
PR curve for fold 6, AUPRC:0.06
PR curve for fold 7, AUPRC:1.00
PR curve for fold 8, AUPRC:0.12
PR curve for fold 9, AUPRC:1.00
PR curve for fold 10, AUPRC:1.00
Mean PR (AUPRC = 0.67 ± nan)

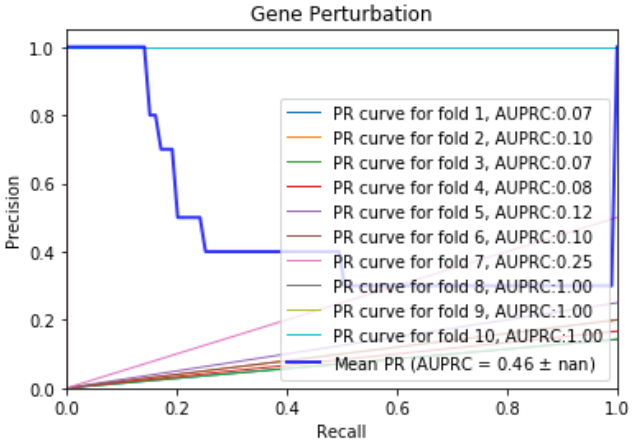

Medium Type
True Positive Rate
False Positive Rate
ROC curve for fold 1, area:1.00
ROC curve for fold 2, area:0.94
ROC curve for fold 3, area:1.00
ROC curve for fold 4, area:1.00
ROC curve for fold 5, area:1.00
ROC curve for fold 6, area:1.00
ROC curve for fold 7, area:1.00
ROC curve for fold 8, area:1.00
ROC curve for fold 9, area:1.00
ROC curve for fold 10, area:1.00
Luck
Mean ROC (AUC = 0.99 ± 0.02)


Medium Type
Precision
Recall
PR curve for fold 1, AUPRC:1.00
PR curve for fold 2, AUPRC:0.25
PR curve for fold 3, AUPRC:1.00
PR curve for fold 4, AUPRC:1.00
PR curve for fold 5, AUPRC:1.00
PR curve for fold 6, AUPRC:1.00
PR curve for fold 7, AUPRC:1.00
PR curve for fold 8, AUPRC:1.00
PR curve for fold 9, AUPRC:1.00
PR curve for fold 10, AUPRC:1.00
Mean PR (AUPRC = 0.95 ± nan)


Environmental Perturbation
True Positive Rate
False Positive Rate
ROC curve for fold 1, area:nan
ROC curve for fold 2, area:0.71
ROC curve for fold 3, area:0.14
ROC curve for fold 4, area:0.71
ROC curve for fold 5, area:0.71
ROC curve for fold 6, area:1.00
ROC curve for fold 7, area:0.71
ROC curve for fold 8, area:1.00
ROC curve for fold 9, area:1.00
ROC curve for fold 10, area:0.00
Luck
Mean ROC (AUC = nan ± nan)

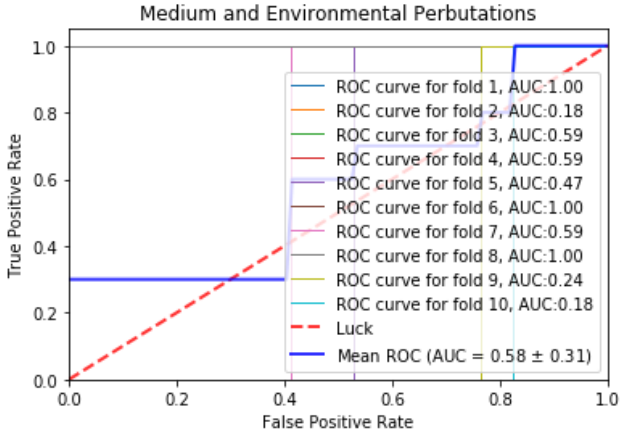

Environmental Perturbation
Precision
Recall
PR curve for fold 1, AUPRC:nan
PR curve for fold 2, AUPRC:0.17
PR curve for fold 3, AUPRC:0.07
PR curve for fold 4, AUPRC:0.17
PR curve for fold 5, AUPRC:0.17
PR curve for fold 6, AUPRC:1.00
PR curve for fold 7, AUPRC:0.17
PR curve for fold 8, AUPRC:1.00
PR curve for fold 9, AUPRC:1.00
PR curve for fold 10, AUPRC:0.06
Mean PR (AUPRC = nan ± nan)





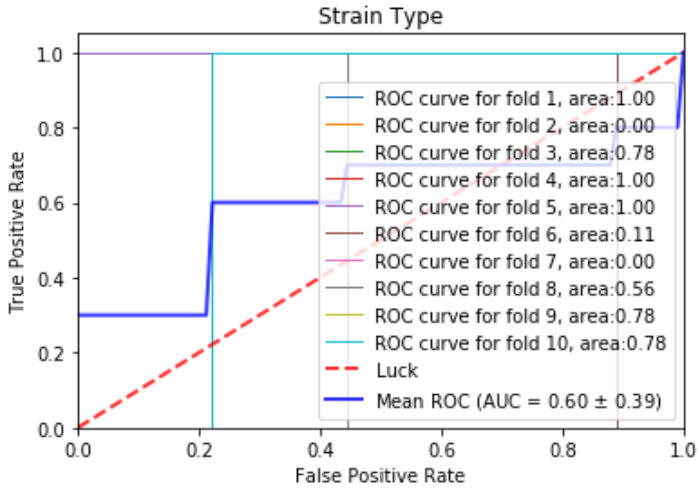
5) Refer to part5.py

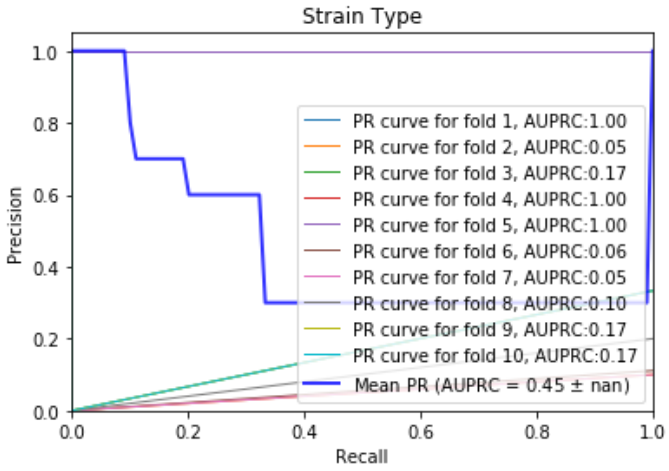
Didn't finish graphing this problem, but I assumed that the composite classifier works better than two separate classifiers since we have more data to work with. Since we want to test whether simultaneous prediction works the same as separate, if they all have similar accuracy in terms of AUC/AUPRC then simultaneous is the same as separate. The null hypothesis is a base of the comparison in which 3 classifications gives equal accuracy. The alternative hypothesis is where simultaneous is worse than separate. Since not all class combinations apply to the composite model, in the null hypothesis we would classify randomly.

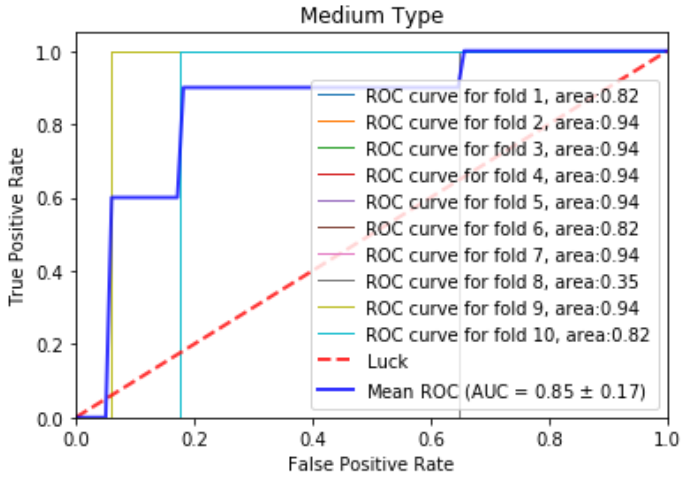


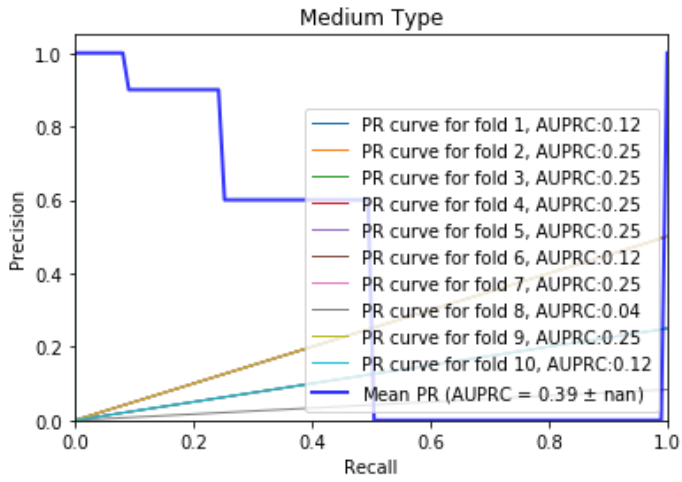
6) Refer to part6_ROC.py and part6_PR.py

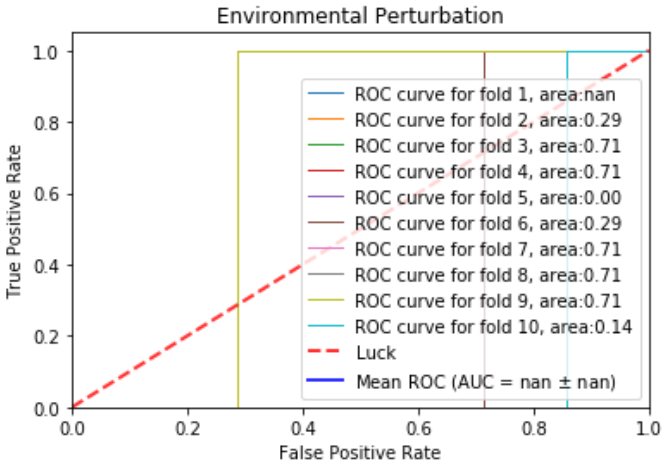
Judging from the AUC/AUPRC values below in the graphs, the PCs do not retain classification performance as well as the SVM classifiers from the previous questions because the AUC/AUPRC values are lower.

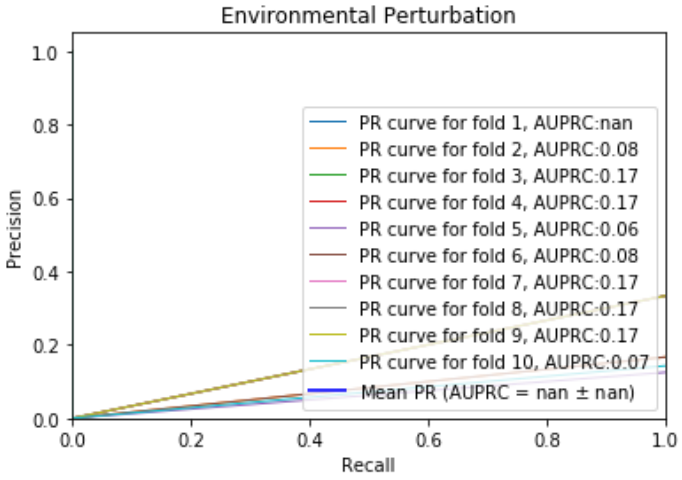


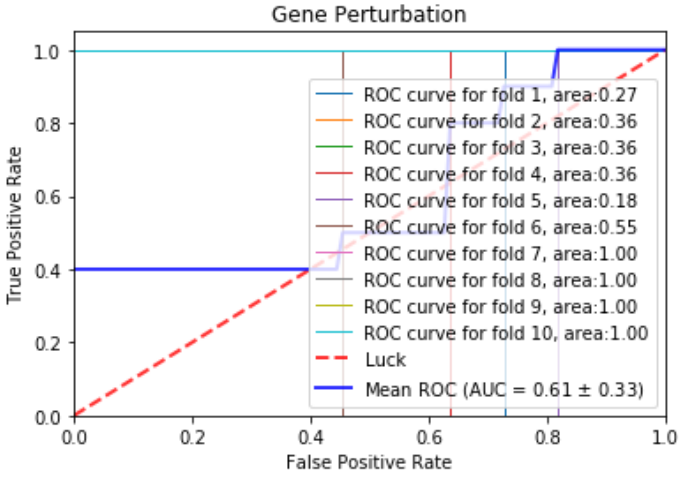


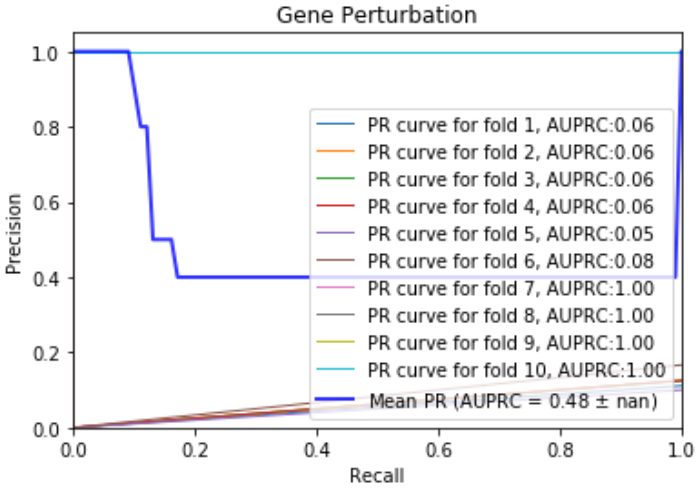





Environmental Perturbation
True Positive Rate
False Positive Rate
ROC curve for fold 1, area:nan
ROC curve for fold 2, area:0.29
ROC curve for fold 3, area:0.71
ROC curve for fold 4, area:0.71
ROC curve for fold 5, area:0.00
ROC curve for fold 6, area:0.29
ROC curve for fold 7, area:0.71
ROC curve for fold 8, area:0.71
ROC curve for fold 9, area:0.71
ROC curve for fold 10, area:0.14
Luck
Mean ROC (AUC = nan ± nan)


Environmental Perturbation
Precision
Recall
PR curve for fold 1, AUPRC:nan
PR curve for fold 2, AUPRC:0.08
PR curve for fold 3, AUPRC:0.17
PR curve for fold 4, AUPRC:0.17
PR curve for fold 5, AUPRC:0.06
PR curve for fold 6, AUPRC:0.08
PR curve for fold 7, AUPRC:0.17
PR curve for fold 8, AUPRC:0.17
PR curve for fold 9, AUPRC:0.17
PR curve for fold 10, AUPRC:0.07
Mean PR (AUPRC = nan ± nan)


Gene Perturbation
True Positive Rate
False Positive Rate
ROC curve for fold 1, area:0.27
ROC curve for fold 2, area:0.36
ROC curve for fold 3, area:0.36
ROC curve for fold 4, area:0.36
ROC curve for fold 5, area:0.18
ROC curve for fold 6, area:0.55
ROC curve for fold 7, area:1.00
ROC curve for fold 8, area:1.00
ROC curve for fold 9, area:1.00
ROC curve for fold 10, area:1.00
Luck
Mean ROC (AUC = 0.61 ± 0.33)


Gene Perturbation
Precision
Recall
PR curve for fold 1, AUPRC:0.06
PR curve for fold 2, AUPRC:0.06
PR curve for fold 3, AUPRC:0.06
PR curve for fold 4, AUPRC:0.06
PR curve for fold 5, AUPRC:0.05
PR curve for fold 6, AUPRC:0.08
PR curve for fold 7, AUPRC:1.00
PR curve for fold 8, AUPRC:1.00
PR curve for fold 9, AUPRC:1.00
PR curve for fold 10, AUPRC:1.00
Mean PR (AUPRC = 0.48 ± nan)