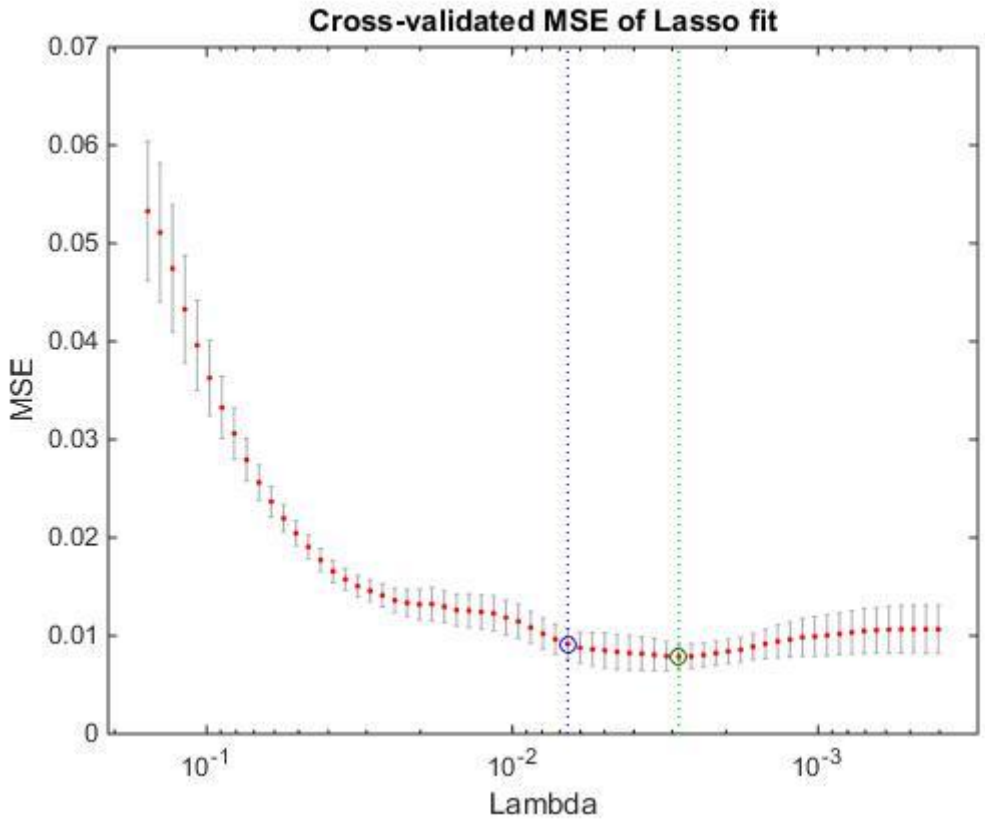
ECS 171

Q1.

I used lasso regularized regression, it helped me to fit regression coefficients for a set of regularization coefficients. Use lassoPlot to plot the graph to see the coross-validated fits.

since lasso output include the fit information, so that we could see there are 1*65 lambda, the optimal constrained parameter value is 0.00659. According to FitInfo I found the DF(Number of nonzero coefficients in B for each value of Lambda, a 1-by-L vector) is 1*65, and 76 nonzero coefficients. 10-fold cross-validation is MSE is 0.012305.



Q2.

The iterations as argument for specified number of interactions of bootstrapping, after resampled the dataset numbers of time. We could get the prediction Y, then

I could calculate the mean and STD, and then get the confidence interval for Y.

Q2

computes the 95% bootstrap confidence interval of the statistic computed by the function capable, I sampled the dataset 100 times, and built 100 models for the datasets. If capable returns a scalar, ci is a vector containing the lower and upper bounds of the confidence interval. If capable returns a vector of length *m*, ci is an array of size 2-by-*m*, whereci(1,:) are lower bounds and ci(2,:) are upper bounds. If capable returns an array of size*m*-by-*n*-by-*p*-by-..., ci is an array of size 2-by-*m*-by-*n*-by-*p*-by-..., where ci(1,:,:,:,...) is an array of lower bounds and ci(2,:,:,:,...) is an array of upper bounds.

ci =

0.0420

0.0968

Q3

Mean expression value : 0.3936

Value =mean(y)+ [mean (x) – mean(x)]*rate = 0.3936

Q4

I created four classifiers to categorize the strain type, medium type, environmental and gene perturbation. Each one of them have more than 10 different factors, so I run SVM for each one of these different factors. Each of them them is built by combining several binary classifier. I had about 4096

features, it reduced to 76 after feature selection. In each binary model, I use 10 fold CV to get the prediction for each sample. Positive score always means the sample should be classified, otherwise not.

In order to get ROC and PR, I have the score matrix from previously , true class and thresholds. Then we could map scores into interval (0,1). After that we could plot ROC and PRC.

In order to get AUC and AUPRC value, I filled area 2-D plot.

For strain, number of feature is 76, 10 fold CV rate : 1.2064 0.3020, AUC: 0.8346 0.8757 AUPRC: 0.4573 0.3101

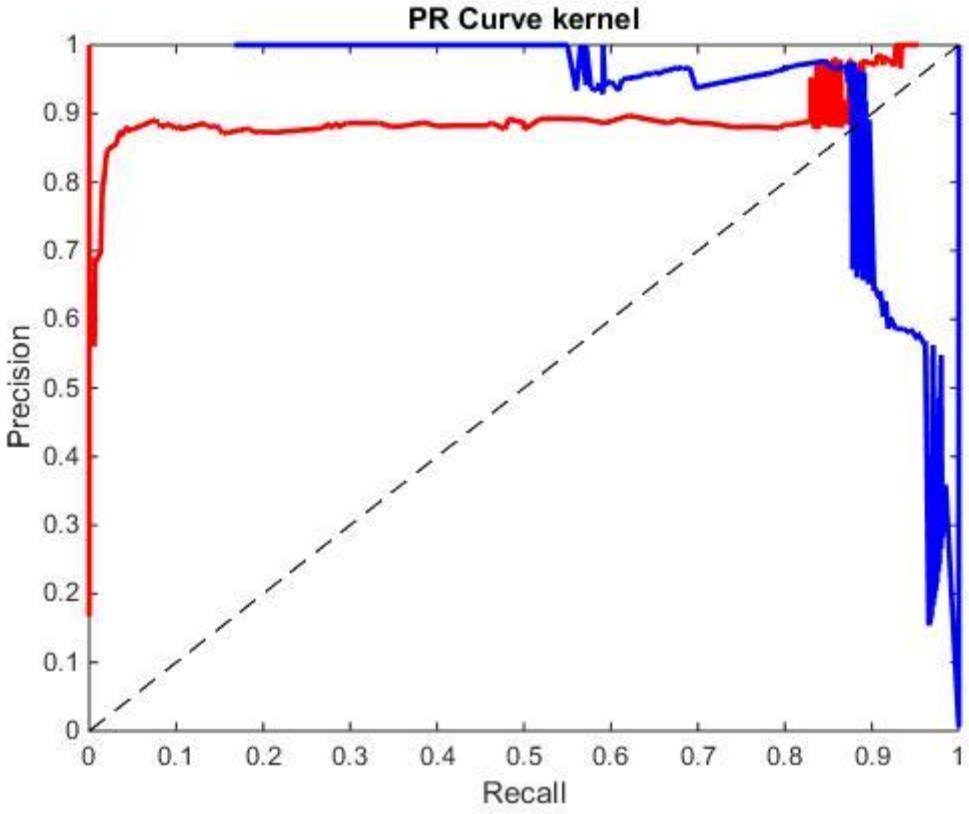
For medium, number of feature is 76, 10 fold CV rate : 0.2403 0.3114 AUC: 0.7399 0.8641 AUPRC: 0.5967 0.5812

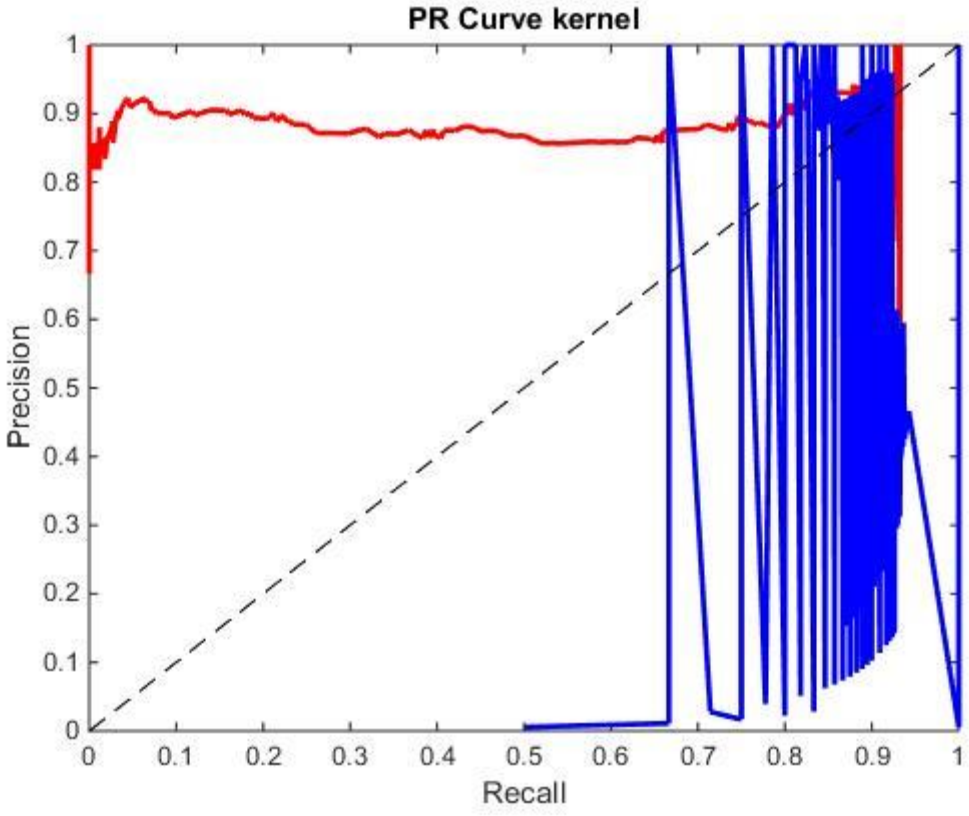
For Stress, number of feature is 76, 10 fold CV rate : 0.2064 0.2010 AUC: 0.8195 0.8636 AUPRC: 0.6751 0.6981

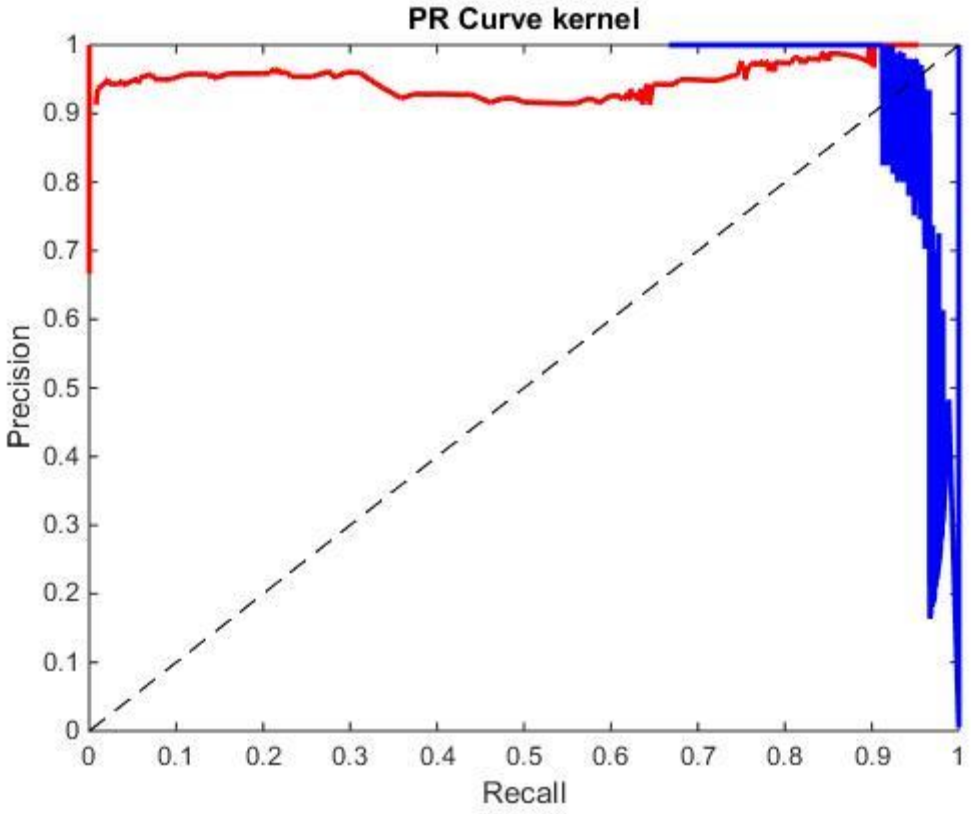
For GenePerturbed , number of feature is 76, 10 fold CV rate : 0.1264 0.1310 AUC: 0.8055 0.8393 AUPRC: 0.5866 0.1111

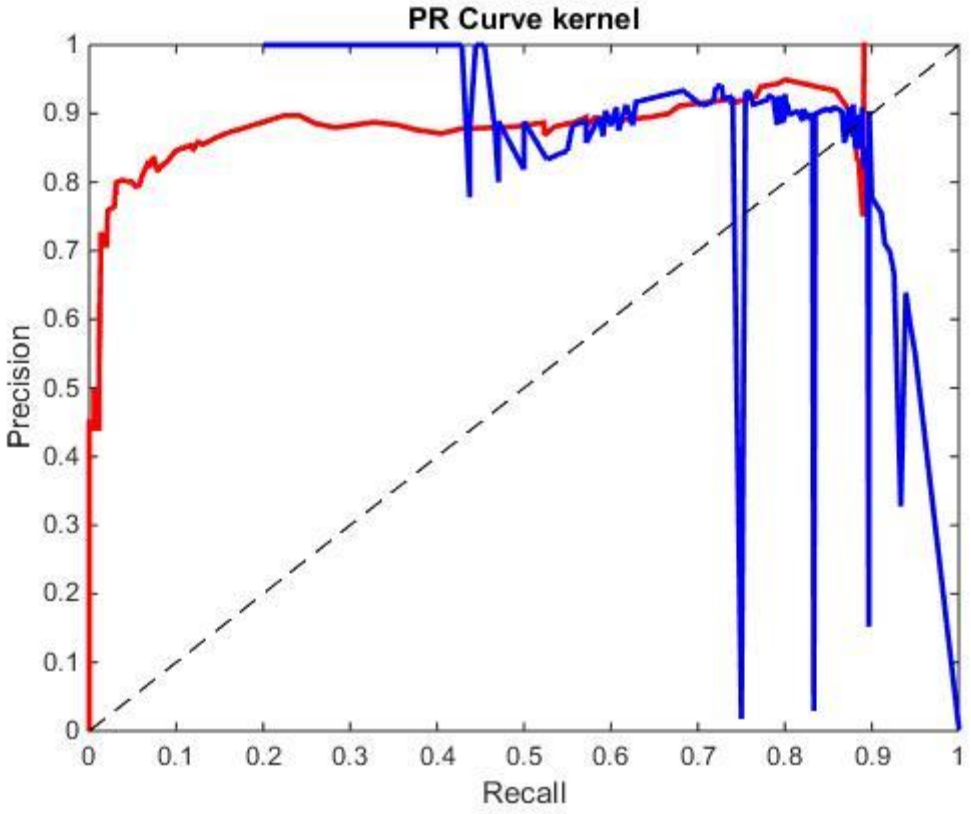
should be 180 classes but the composite model only 26 of them. Which means if class combination out of the 26, it will certainly misclassified. So composite model also has disadvantage side.

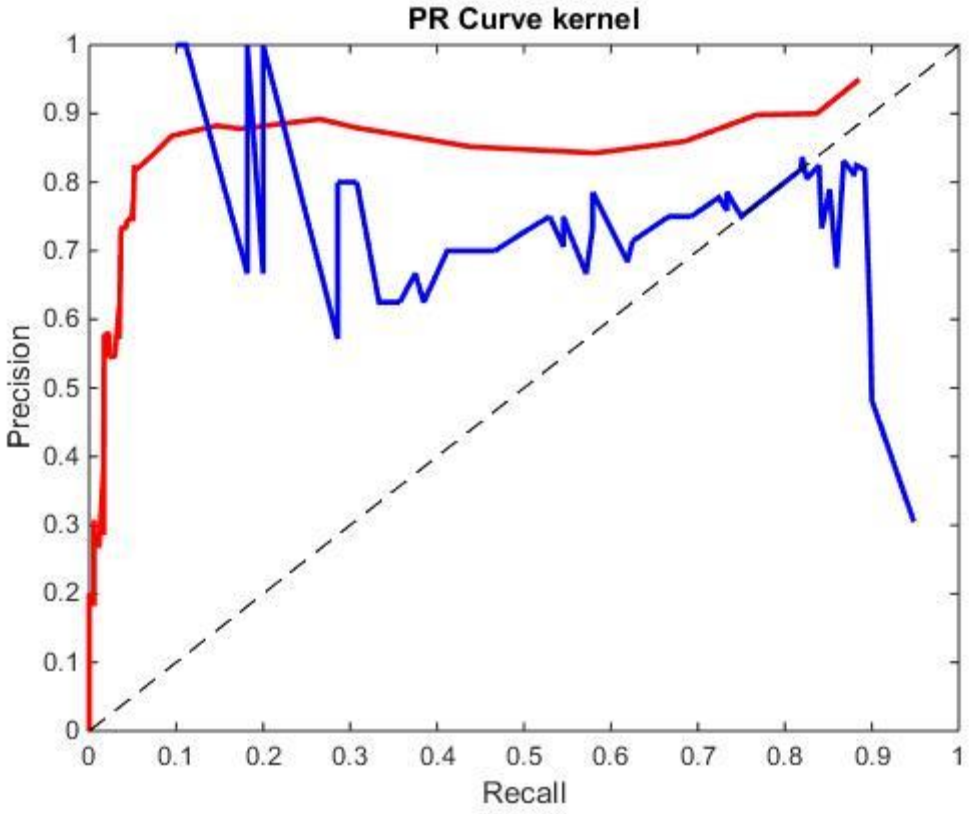
In the null hypothesis, we could classify randomly. The base line should be 0.0385.

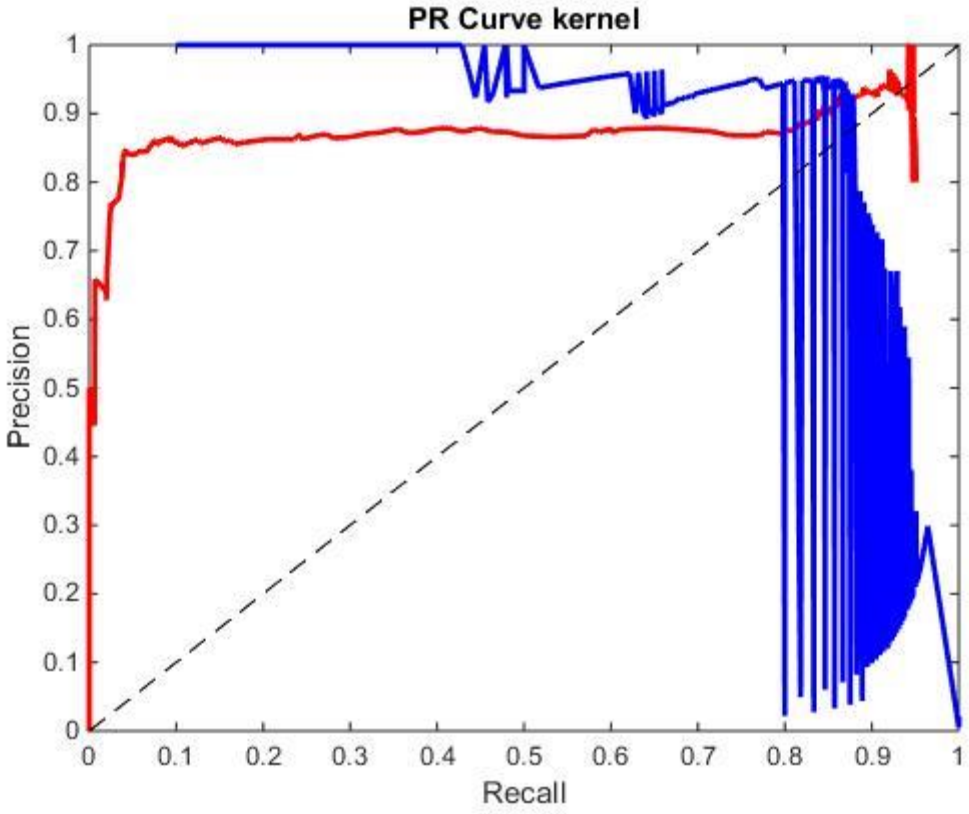

PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

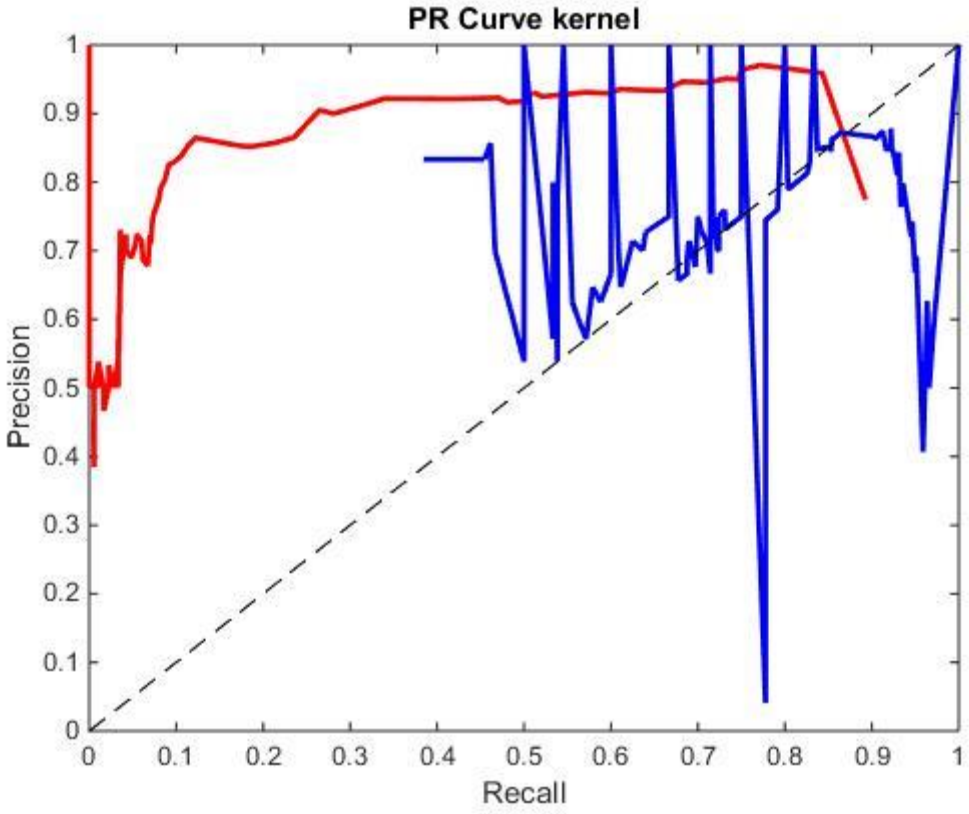

PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

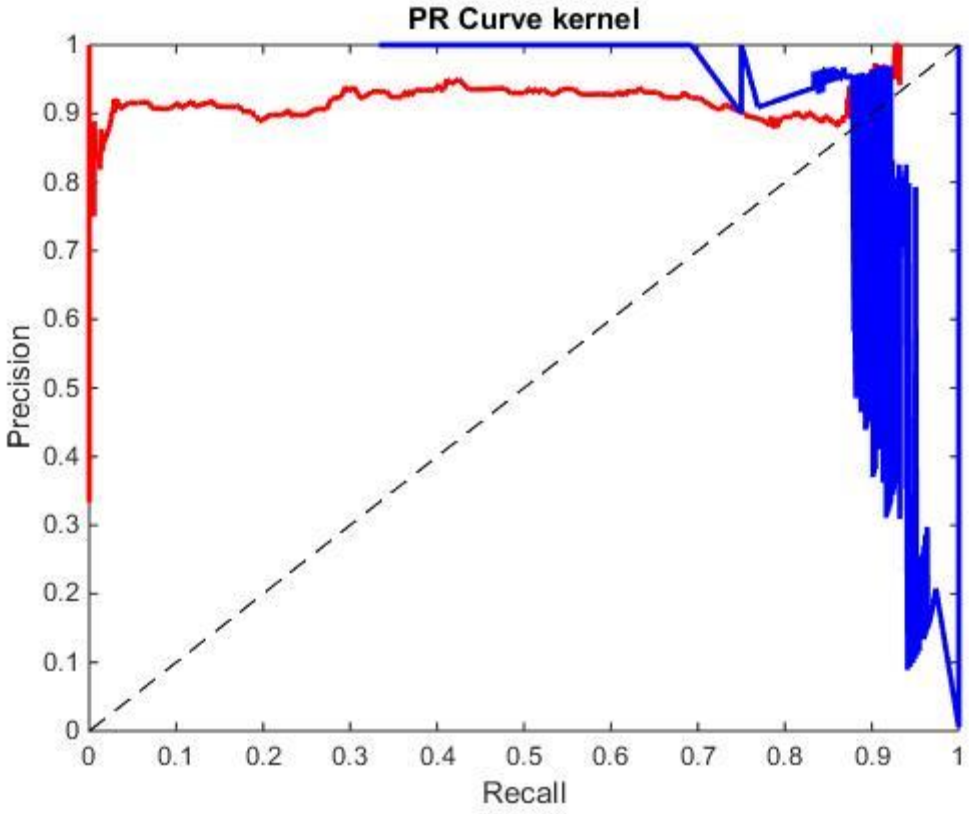

PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


PR Curve kernel
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

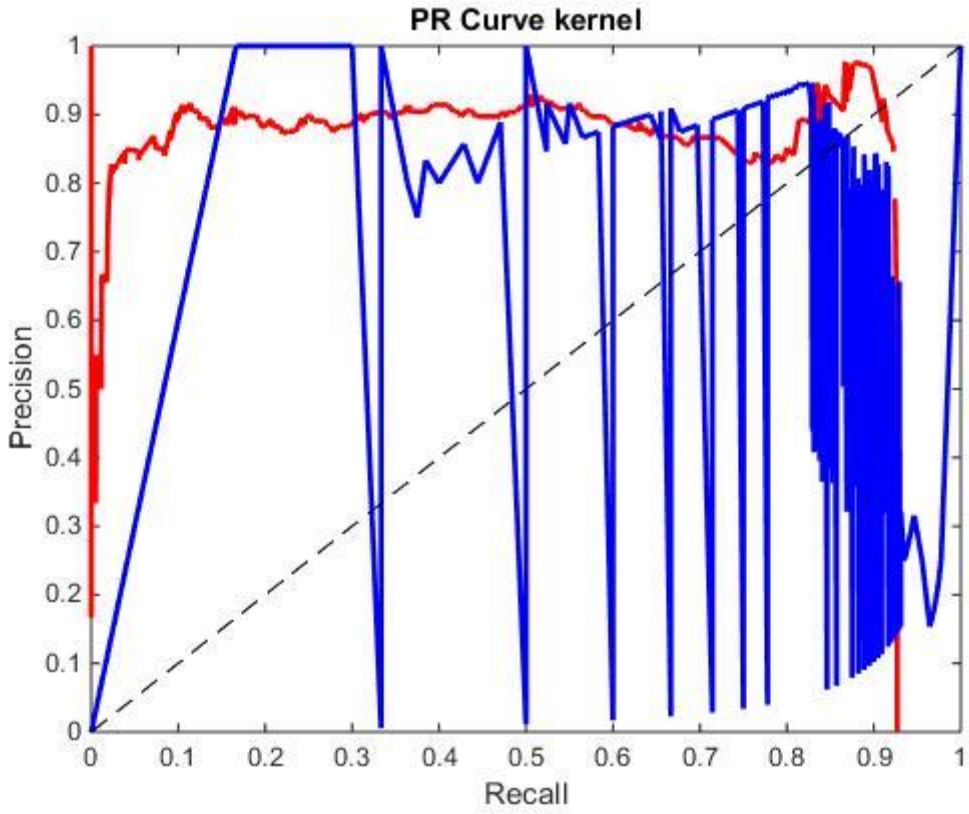
Q5

For composite model ,

AUC: 0.7760

AUPRC: 0.6490

The misclassification rate is 0.3508. it's better than two individual classifiers. Since not all class combinations are apply to composite model. in the null hypothesis, I use randomly classifier, so the separate model base line performance is 0.0056



Q6

For composite model,

AUC: 0.0831

AUPRC: 0.7143

It's very obviously to see the AUC in PC composite and PC medium model

are lower than the SVM models. And perturbation model was not effect a lot.

Conclusion, I don’t think PCs retain most of the classification performance while reducing the dimensionality.