Homework 3
ECS171

Atharva Chalke

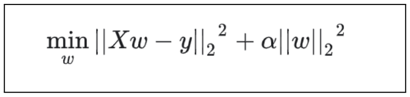
December 5, 2018

Problem 1

Create a predictor of the bacterial growth attribute by using only the expression of the genes as attributes. Not all genes are informative for this task, so use a regularized regression technique (e.g. lasso, ridge or elastic net) and explain what it does (we have not covered the specifics of each method, so you have to do some reading). Which one is the optimal constrained parameter value (usually denoted by)? Report the number of features that have non-zero coefficients and the 10-fold cross-validation generalization error of the technique. [20pt]

Answer

Code: Part1.py
To prevent the model from overfitting, we use regularization to add a penalty to weight updates and decreases the complexity of the model.This is important in our case since we have 192 samples, and are using 4496 features.
Ridge Regression:
Model tries to minimize this loss function in code implemented.



$$\min_{w} ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$$

1. Adds a penalty of square of magnitude of weights.
2. The minimization objective in our case is RSS + Regularization term.
3. If $\alpha = 0$, this is linear-regression.
4. If $\alpha = \infty$, the weights would be 0 since this is the only way to make the loss function minimum as the penalty would be high from α.
5. If $0 < \alpha < \infty$, then the magnitude of α defines the weight-age given to weights. Non-zero values would define how much weight is given to weights.

Optimal Coefficient: 1.5
Features with Non-zero Coefficients: 4434
Generalization error with 10 folds cross validation using MSE:0.008284914606722284

Problem 2

Extend your predictor to report the confidence interval of the prediction by using the bootstrapping method. Clearly state the methodology and your assumptions. [10pt]

Answer

Code: Part2.py
We use resampling with replacement in our implementation of bootstrapping, assuming the random data is representative of our bacterial growth attribute in the real world. 100 iterations are used with a resample size of 174.(approximately 90%). A new model is trained on the resampled data, and a prediction is made on the value you want to report the confidence interval on. All of this is appended to a list, that is later used to find the CI.
We try to find the 95% confidence interval, by taking off the tail on both ends of size (100-95)/2=5/2 = 2.5.

Problem 3

Use your bootstrap model from 2 to find the confidence interval of predicted growth for a bacterium whose genes are expressed exactly at the mean expression value. (Note: for each gene, there is a corresponding mean expression value) [5pt]

Answer

Code: Part3.py
95.0 confidence interval 0.3929 and 0.4087

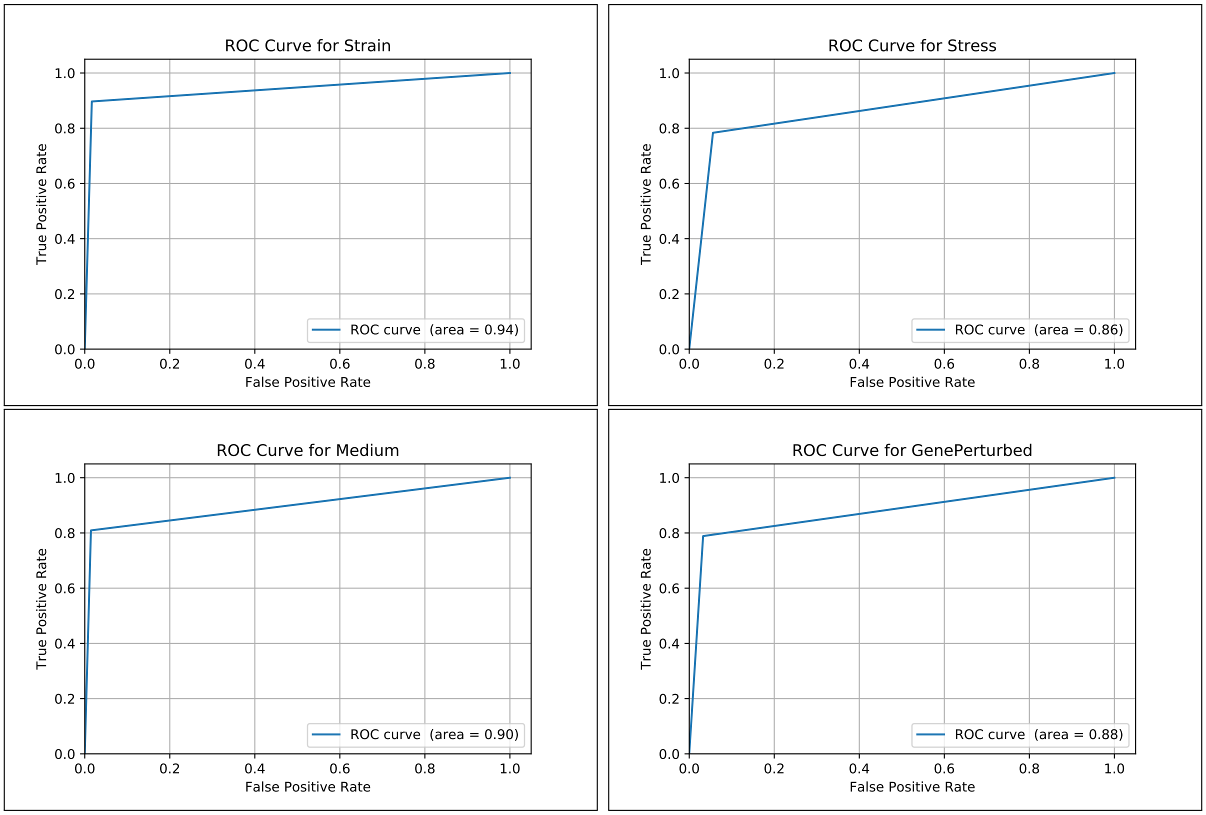
Problem 4

Create four separate SVM classifiers to categorize the strain type, medium type, environmental and gene perturbation, given all the gene transcriptional profiles. The classifier should select as features a small subset of the genes, either by performing feature selection (wrapper method) or by using only the non-zero weighted features from the regularized regression technique of the first aim. For each classifier (4 total) report the number of features and the classification performance though 10-fold cross-validation by plotting the ROC and PR curves and reporting the AUC/AUPRC values.[20pt]

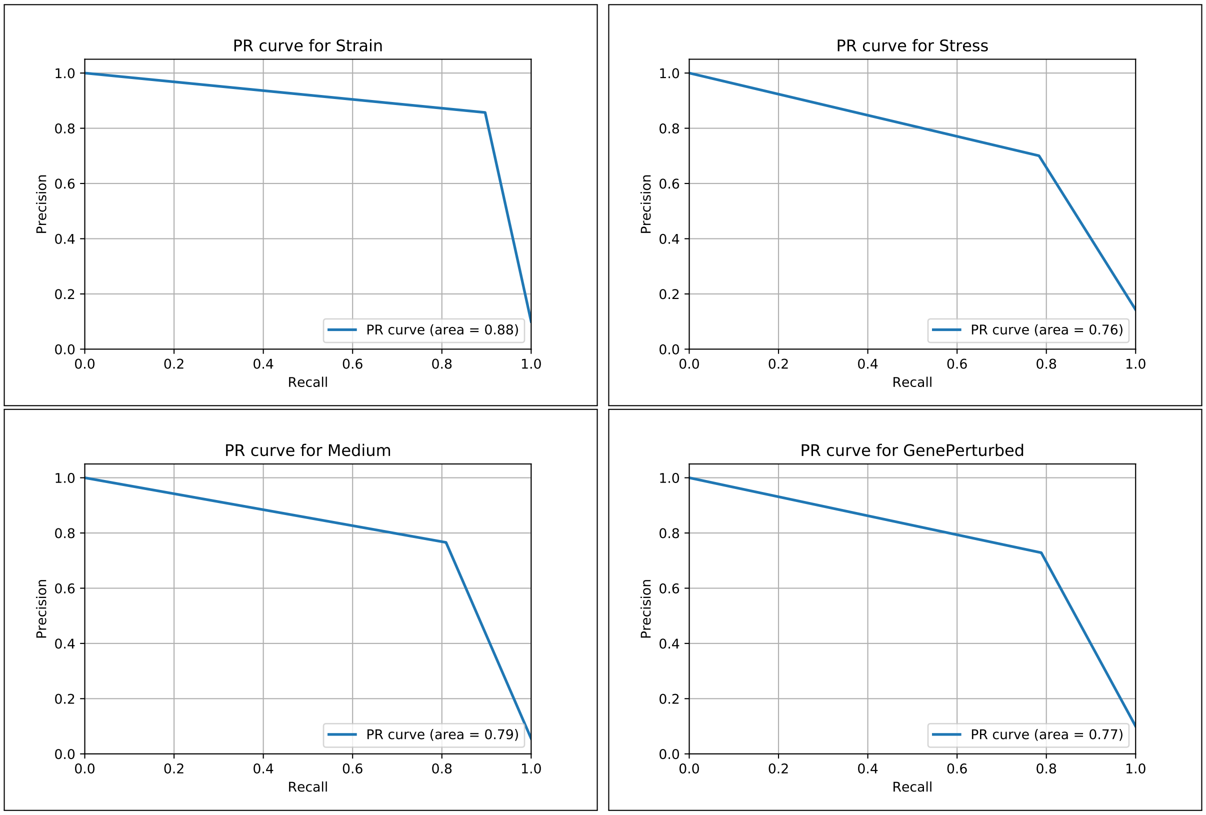
Answer

Code: Part4.py
For all SVMS, we used the non-zero features obtained from Ridge Regression.
We have 4434 features for all SVMs.

ROC Curves with AUC



Precision Recall Curves with AUPRC

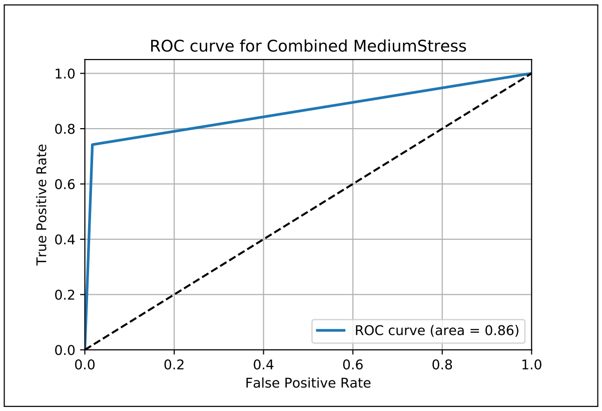


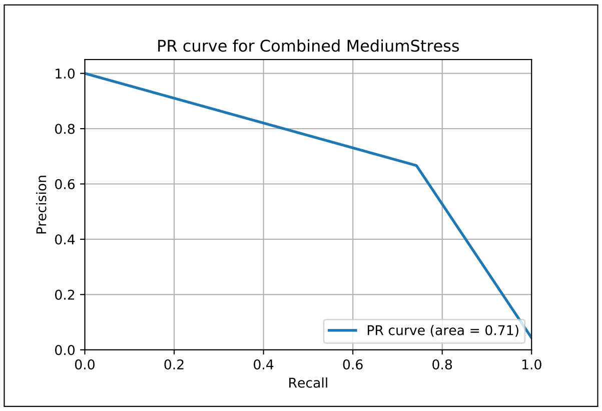
Problem 5

Create one composite SVM classifier to simultaneously predict medium and environmental perturbations and report the 10-fold cross-validation AUC/AUPRC value. Does this classifier perform better or worse than the two individual classifiers together for these predictions? That is, are we better off building one composite or two separate classifiers to simultaneously predict these two features? What is the baseline prediction performance(null hypothesis)? [15pt]

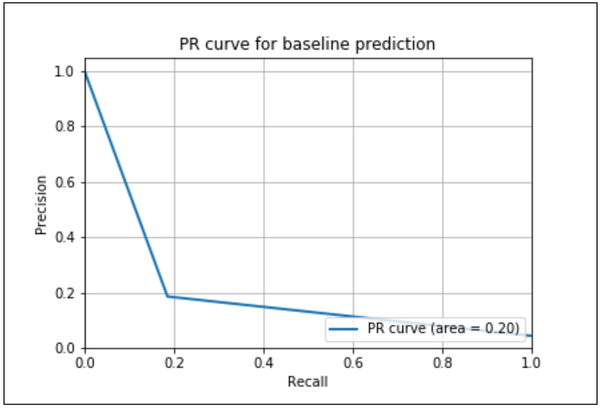
Answer

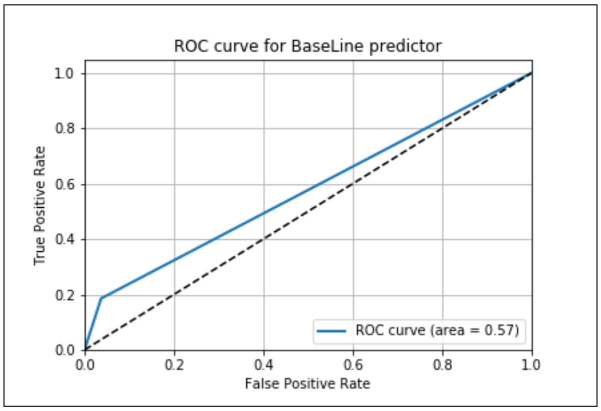
Code: Part5.py





The ROC AUC under the combined classifier is the same as the ROC AUC under the Stress Graph; however, it is lower than the ROC AUC for just the Medium classifier.
The PR AUC under the combined classifier is lower than the PR AUC for the individual Medium and Stress classifier.
Therefore, we are better off building two individual classifiers, instead of a combined classifier since the AUC shows that the classification performance becomes worse if we build a combined classifier.





For baseline predictions, we predicted the most occuring combination every time. As we can see, our classifiers performance is much better than that our null hypothesis.

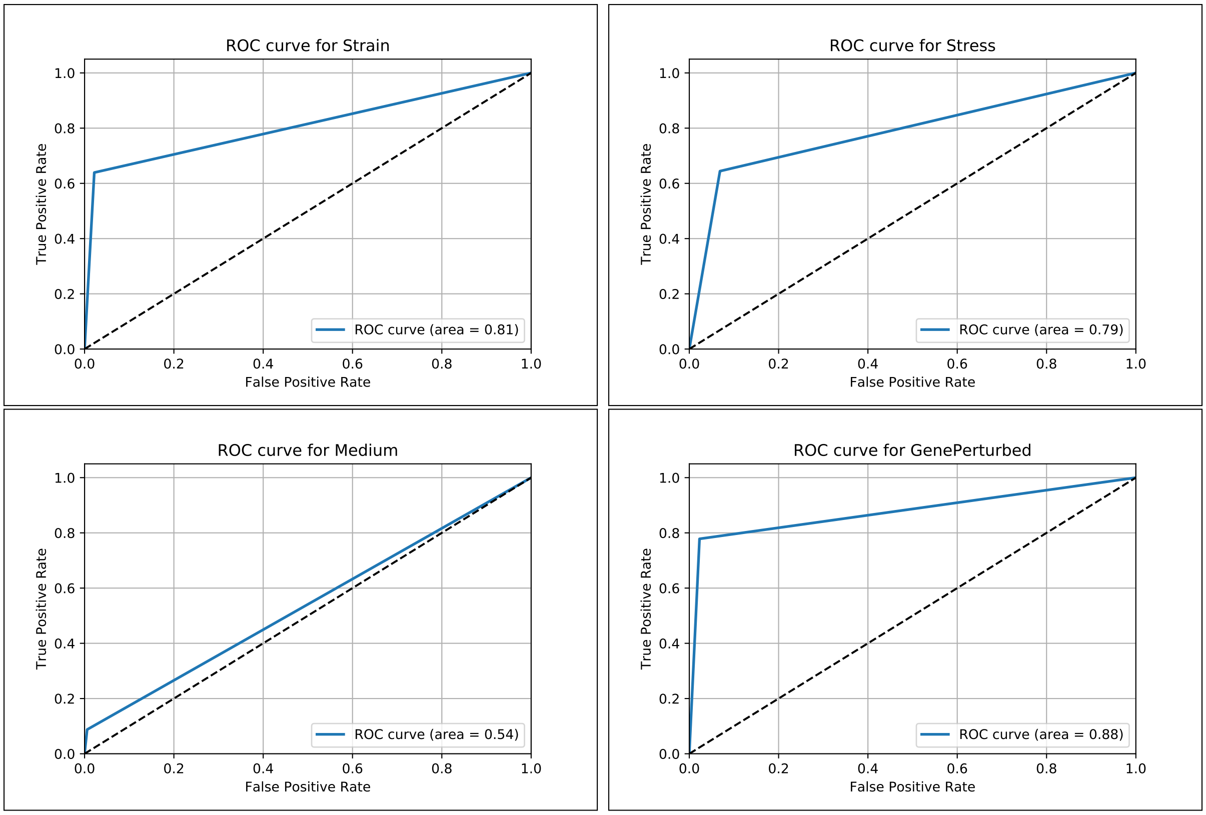
Problem 6

Perform Principal Component Analysis, keeping only the 3 Principal Components (PCs)as features for the SVM classifier (no other features except of those three). Report the10-fold cross-validation AUC/AUPRC value and plot the ROC/PR curves on the same plot as before. Do the PCs retain most of the classification performance while reducing the dimensionality? [10pt]

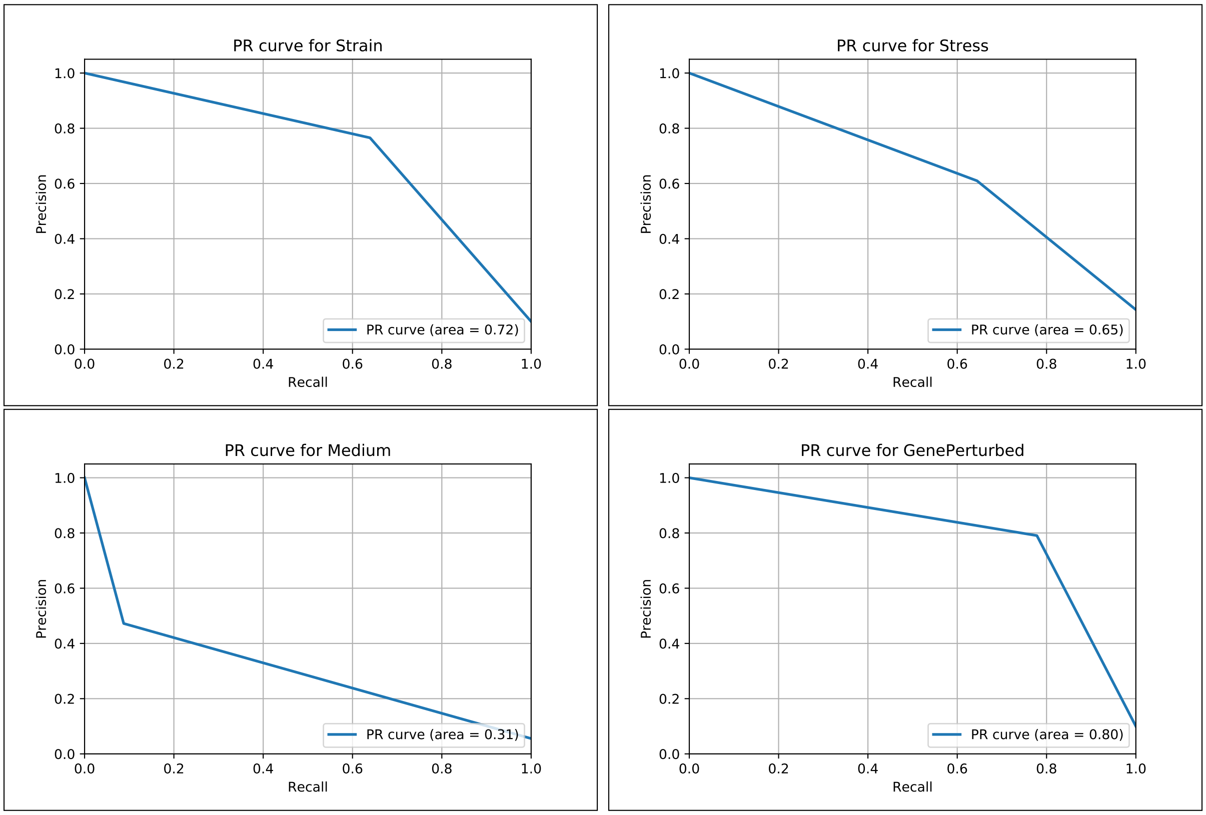
Answer

Code: Part6.py

PCA ROC Curves with AUC



PCA Precision Recall Curves with AUPRC



PCA does retain a lot of information; however, it doesn't retain all of the information as can be seen from the ROC and PR plots since the classification performance decreases in most cases as can be seen by the decrease in AUC for PR and ROC curves.