Report hw3
Name: Steven Gong

Files:

1. main.m  It is the main file I did LASSO here and feature selection as well as running other files
2. svm.m . I did 10-fold cross validation here.
3. trainingsvm.m. This is for training the objective function, I implement the code with SMO(Sequential minimal optimization) algorithm. I will train 50000 times or until converge.
4. area.m I implement the trapezoid area estimation to get AUC.
5. kernel.m, kernel function, however, I choose the dot product
6. ecs171.dataset.xlsx
7. medium.txt , I convert stirngs to numeric values
8. GP.txt
9. strain.txt
10. stress.txt

To run the code just run from main. Notice: I did not use any package. So I should get **20 points for bonus points ☺**
It might take a while though. It might take up 4 hours, because my code is not efficient. It's mainly because of the svm part. However, the percentage of correct is pretty high.

I divided the data into 10 folds, each contains 19 data
Please notice: since there are a lot of graphs in my methods(200), so the computer might take a while to get all the graphs. I plot a ROC PR curve for each classifier and each fold.
Because I do not have time to run through all the categories.

Q1:
I am using LASSO, it will give a constrain like a diamond, and some of the weights of the feature will become 0, due to its shape.; therefore, I have done some feature selection. The probability of being zero is very high.
The update rule is : $w = w(|w|/w - a * l/m) - a *(h(x, w) - y)* x/m$

I split the dataset into 10 pieces, each piece contains 19 sample. Then I do the 10-fold cross validation. The generalization error is obtained through average the error between training set and the testing set.
When I run 500 times I got
$\lambda = 0.1$ generalization error = -0.0887
$\lambda = 1$ generalization error = -0.0220
$\lambda = 3$ generalization error = -0.0190
$\lambda = 10$ generalization error = 0.0075
$\lambda = 60$ generalization error = 0.0619
$\lambda = 100$ generalization error = 0.0329

so the optimal λ is **10**

if the weight is greater than -0.0001 and less than 0.0001, we would regard it as 0.
the number of non-zero w is: **313**
The optimal model is #5, so I choose that as my wmin. (optimal w)

Q2
I split the data into 10 portions. I use 9 of them to train, and 1 as testing set. I get 10 different models, and their corresponding predictions. Then I calculate the standard deviation. Then I can get the confidence interval with 95% confidence interval.
Confidence interval: for 194 predictions:

| | |
|---|---|
| 0.4830 | 0.4853 |
| 0.4852 | 0.4876 |
| 0.4880 | 0.4904 |
| 0.5166 | 0.5194 |
| 0.5106 | 0.5134 |
| 0.5120 | 0.5147 |
| 0.4316 | 0.4333 |
| 0.4340 | 0.4357 |
| 0.4355 | 0.4373 |
| 0.4370 | 0.4387 |
| 0.4394 | 0.4411 |
| 0.4403 | 0.4420 |
| 0.4388 | 0.4405 |
| 0.4388 | 0.4405 |
| 0.4522 | 0.4541 |
| 0.4399 | 0.4414 |
| 0.4547 | 0.4564 |
| 0.4517 | 0.4534 |
| 0.4540 | 0.4559 |
| 0.4515 | 0.4533 |
| 0.3353 | 0.3363 |
| 0.3329 | 0.3361 |
| 0.4614 | 0.4633 |
| 0.3045 | 0.3057 |
| 0.3943 | 0.3962 |
| 0.3934 | 0.3954 |
| 0.3945 | 0.3965 |
| 0.4036 | 0.4056 |
| 0.4056 | 0.4075 |
| 0.4068 | 0.4088 |
| 0.4248 | 0.4270 |
| 0.4261 | 0.4284 |
| 0.4369 | 0.4390 |
| 0.4305 | 0.4325 |

| | |
|---|---|
| 0.4277 | 0.4300 |
| 0.4275 | 0.4297 |
| 0.4673 | 0.4698 |
| 0.4759 | 0.4785 |
| 0.4687 | 0.4712 |
| 0.4723 | 0.4749 |
| 0.4692 | 0.4719 |
| 0.4695 | 0.4722 |
| 0.4814 | 0.4842 |
| 0.3762 | 0.3795 |
| 0.4930 | 0.4958 |
| 0.4927 | 0.4956 |
| 0.4954 | 0.4981 |
| 0.3476 | 0.3487 |
| 0.3249 | 0.3258 |
| 0.3820 | 0.3833 |
| 0.3727 | 0.3739 |
| 0.4298 | 0.4316 |
| 0.4298 | 0.4316 |
| 0.4324 | 0.4343 |
| 0.3617 | 0.3628 |
| 0.3578 | 0.3588 |
| 0.3618 | 0.3628 |
| 0.3877 | 0.3888 |
| 0.4104 | 0.4118 |
| 0.3494 | 0.3503 |
| 0.3498 | 0.3507 |
| 0.3492 | 0.3502 |
| 0.2702 | 0.2712 |
| 0.2819 | 0.2830 |
| 0.2809 | 0.2819 |
| 0.2696 | 0.2706 |
| 0.3976 | 0.3988 |
| 0.3976 | 0.3988 |
| 0.4068 | 0.4083 |
| 0.3661 | 0.3670 |
| 0.3665 | 0.3674 |
| 0.3638 | 0.3647 |
| 0.3605 | 0.3615 |
| 0.3597 | 0.3607 |
| 0.3602 | 0.3612 |
| 0.3779 | 0.3793 |
| 0.3836 | 0.3848 |
| 0.3840 | 0.3854 |
| 0.3773 | 0.3786 |
| 0.3801 | 0.3814 |

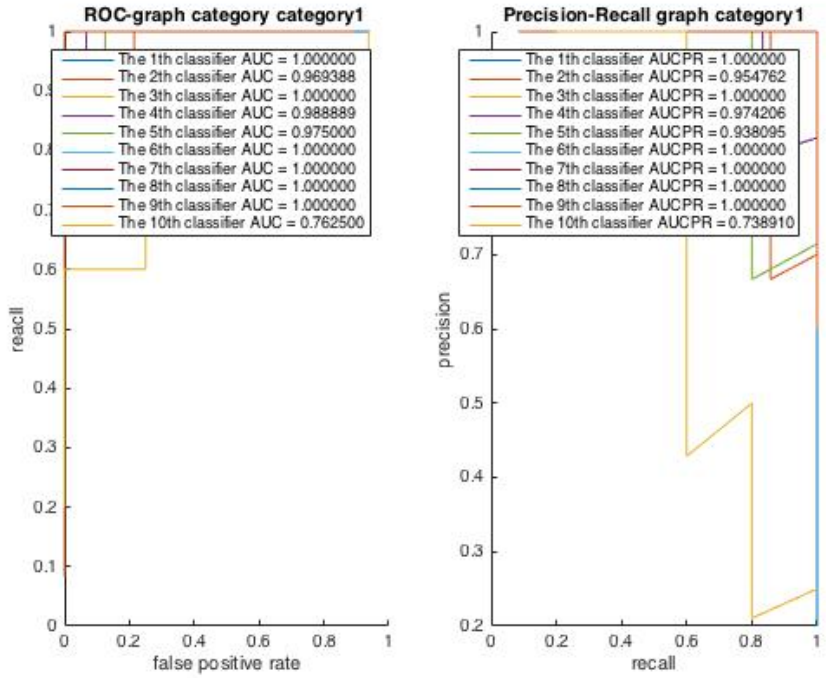
| | |
|---|---|
| 0.3787 | 0.3803 |
| 0.2489 | 0.2508 |
| 0.2496 | 0.2515 |
| 0.2576 | 0.2594 |
| 0.2344 | 0.2363 |
| 0.2421 | 0.2441 |
| 0.2475 | 0.2493 |
| 0.2632 | 0.2651 |
| 0.2412 | 0.2431 |
| 0.2470 | 0.2488 |
| 0.2687 | 0.2705 |
| 0.2388 | 0.2408 |
| 0.2436 | 0.2455 |
| 0.2251 | 0.2272 |
| 0.7632 | 0.7702 |
| 0.7785 | 0.7859 |
| 0.7379 | 0.7457 |
| 0.7581 | 0.7650 |
| 0.7626 | 0.7697 |
| 0.4680 | 0.4699 |
| 0.4808 | 0.4830 |
| 0.4762 | 0.4785 |
| 0.4708 | 0.4731 |
| 0.4653 | 0.4670 |
| 0.4690 | 0.4713 |
| 0.2588 | 0.2605 |
| 0.2644 | 0.2662 |
| 0.2470 | 0.2489 |
| 0.2305 | 0.2322 |
| 0.2462 | 0.2481 |
| 0.2470 | 0.2488 |
| 0.2315 | 0.2335 |
| 0.2341 | 0.2360 |
| 0.2653 | 0.2671 |
| 0.2238 | 0.2258 |
| 0.2268 | 0.2288 |
| 0.2709 | 0.2728 |
| 0.3605 | 0.3616 |
| 0.3540 | 0.3551 |
| 0.2746 | 0.2771 |
| 0.3594 | 0.3605 |
| 0.3661 | 0.3671 |
| 0.2763 | 0.2788 |
| 0.3639 | 0.3649 |
| 0.3644 | 0.3655 |
| 0.3635 | 0.3646 |

| | |
|---|---|
| 0.3586 | 0.3596 |
| 0.3643 | 0.3654 |
| 0.2856 | 0.2881 |
| 0.3076 | 0.3102 |
| 0.3227 | 0.3263 |
| 0.3087 | 0.3115 |
| 0.3490 | 0.3549 |
| 0.2977 | 0.3010 |
| 0.3009 | 0.3040 |
| 0.2938 | 0.2968 |
| 0.2888 | 0.2915 |
| 0.3605 | 0.3660 |
| 0.3447 | 0.3472 |
| 0.3450 | 0.3458 |
| 0.4665 | 0.4685 |
| 0.3475 | 0.3513 |
| 0.3876 | 0.3884 |
| 0.3864 | 0.3872 |
| 0.3871 | 0.3878 |
| 0.3847 | 0.3855 |
| 0.3846 | 0.3854 |
| 0.3823 | 0.3831 |
| 0.3821 | 0.3829 |
| 0.3847 | 0.3855 |
| 0.3834 | 0.3842 |
| 0.3849 | 0.3857 |
| 0.3826 | 0.3834 |
| 0.3818 | 0.3826 |
| 0.3858 | 0.3866 |
| 0.3873 | 0.3881 |
| 0.3944 | 0.3953 |
| 0.3583 | 0.3593 |
| 0.4604 | 0.4631 |
| 0.4590 | 0.4609 |
| 0.4626 | 0.4646 |
| 0.2947 | 0.2958 |
| 0.2957 | 0.2969 |
| 0.2969 | 0.2980 |
| 0.2955 | 0.2965 |
| 0.2576 | 0.2589 |
| 0.2930 | 0.2940 |
| 0.2672 | 0.2682 |
| 0.2994 | 0.3005 |
| 0.3735 | 0.3743 |
| 0.2856 | 0.2865 |
| 0.4563 | 0.4580 |

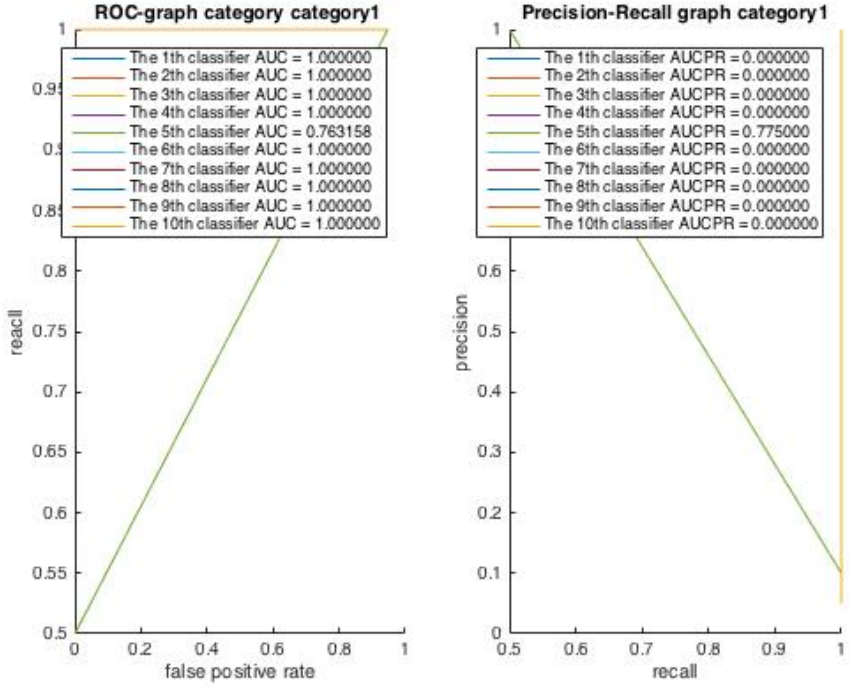
| | |
|---|---|
| 0.2861 | 0.2871 |
| 0.2977 | 0.2985 |
| 0.4787 | 0.4804 |
| 0.4228 | 0.4240 |
| 0.3927 | 0.3938 |
| 0.4569 | 0.4590 |
| 0.4568 | 0.4589 |
| 0.3348 | 0.3363 |
| 0.3331 | 0.3345 |
| 0.3329 | 0.3344 |
| 0.3220 | 0.3233 |
| 0.3234 | 0.3246 |
| 0.3356 | 0.3371 |
| 0.3340 | 0.3354 |
| 0.3195 | 0.3206 |
| 0.3338 | 0.3352 |
| 0.3242 | 0.3255 |
| 0.3215 | 0.3227 |
| 0.3348 | 0.3363 |
| 0.3323 | 0.3338 |
| 0.3319 | 0.3334 |
| 0.3321 | 0.3337 |

Q3:
I select model 5(out of 10 folds) as my model, because its MSE for testing set is the smallest.
predictionQ3 = q3 * wmin + biasmin = 0.3725

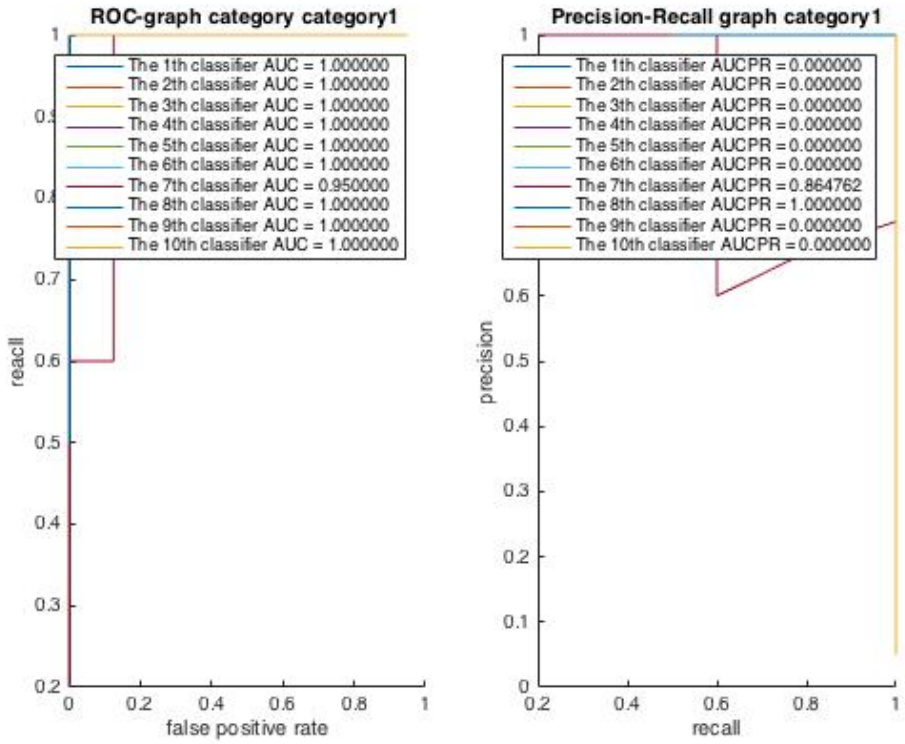
Q4
I use **313** features selected by Q1 through LASSO
Medium:


ROC-graph category category1
The 1th classifier AUC = 1.000000
The 2th classifier AUC = 0.969388
The 3th classifier AUC = 1.000000
The 4th classifier AUC = 0.988889
The 5th classifier AUC = 0.975000
The 6th classifier AUC = 1.000000
The 7th classifier AUC = 1.000000
The 8th classifier AUC = 1.000000
The 9th classifier AUC = 1.000000
The 10th classifier AUC = 0.762500
reacll
false positive rate
Precision-Recall graph category1
The 1th classifier AUCPR = 1.000000
The 2th classifier AUCPR = 0.954762
The 3th classifier AUCPR = 1.000000
The 4th classifier AUCPR = 0.974206
The 5th classifier AUCPR = 0.938095
The 6th classifier AUCPR = 1.000000
The 7th classifier AUCPR = 1.000000
The 8th classifier AUCPR = 1.000000
The 9th classifier AUCPR = 1.000000
The 10th classifier AUCPR = 0.738910
precision
recall

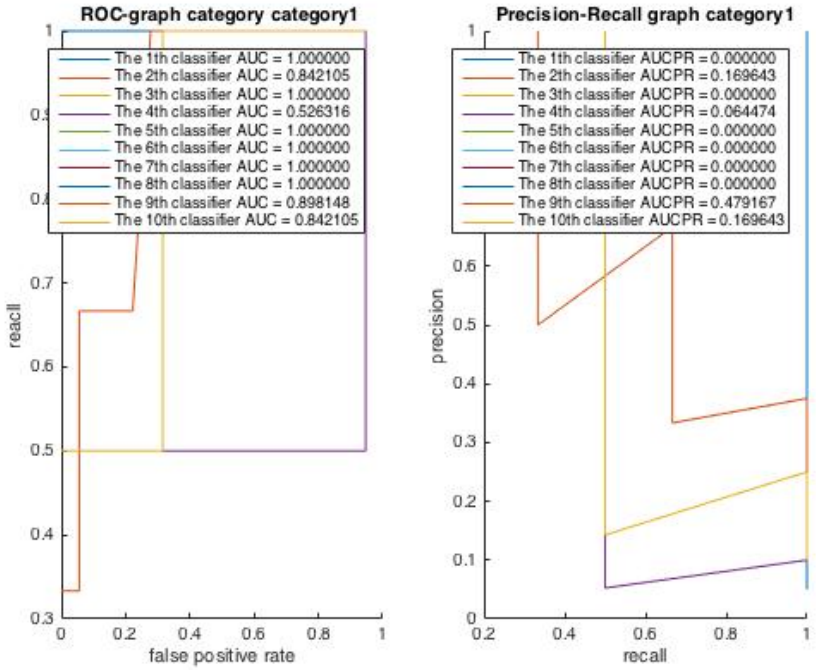
Strain type:


ROC-graph category category1
The 1th classifier AUC = 1.000000
The 2th classifier AUC = 1.000000
The 3th classifier AUC = 1.000000
The 4th classifier AUC = 1.000000
The 5th classifier AUC = 0.763158
The 6th classifier AUC = 1.000000
The 7th classifier AUC = 1.000000
The 8th classifier AUC = 1.000000
The 9th classifier AUC = 1.000000
The 10th classifier AUC = 1.000000
reacll
false positive rate
Precision-Recall graph category1
The 1th classifier AUCPR = 0.000000
The 2th classifier AUCPR = 0.000000
The 3th classifier AUCPR = 0.000000
The 4th classifier AUCPR = 0.000000
The 5th classifier AUCPR = 0.775000
The 6th classifier AUCPR = 0.000000
The 7th classifier AUCPR = 0.000000
The 8th classifier AUCPR = 0.000000
The 9th classifier AUCPR = 0.000000
The 10th classifier AUCPR = 0.000000
precision
recall

GP:

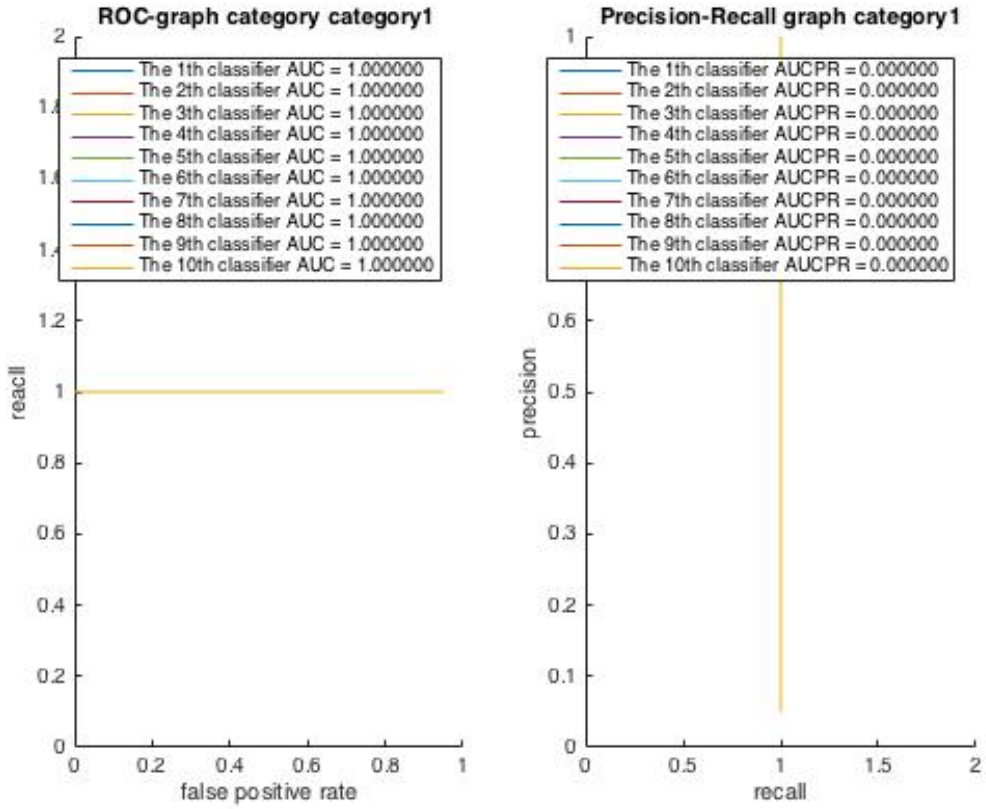

ROC-graph category category1
The 1th classifier AUC = 1.000000
The 2th classifier AUC = 1.000000
The 3th classifier AUC = 1.000000
The 4th classifier AUC = 1.000000
The 5th classifier AUC = 1.000000
The 6th classifier AUC = 1.000000
The 7th classifier AUC = 0.950000
The 8th classifier AUC = 1.000000
The 9th classifier AUC = 1.000000
The 10th classifier AUC = 1.000000
reacll
false positive rate
Precision-Recall graph category1
The 1th classifier AUCPR = 0.000000
The 2th classifier AUCPR = 0.000000
The 3th classifier AUCPR = 0.000000
The 4th classifier AUCPR = 0.000000
The 5th classifier AUCPR = 0.000000
The 6th classifier AUCPR = 0.000000
The 7th classifier AUCPR = 0.864762
The 8th classifier AUCPR = 1.000000
The 9th classifier AUCPR = 0.000000
The 10th classifier AUCPR = 0.000000
precision
recall

Stress:


ROC-graph category category1
The 1th classifier AUC = 1.000000
The 2th classifier AUC = 0.842105
The 3th classifier AUC = 1.000000
The 4th classifier AUC = 0.526316
The 5th classifier AUC = 1.000000
The 6th classifier AUC = 1.000000
The 7th classifier AUC = 1.000000
The 8th classifier AUC = 1.000000
The 9th classifier AUC = 0.898148
The 10th classifier AUC = 0.842105
reacll
false positive rate
Precision-Recall graph category1
The 1th classifier AUCPR = 0.000000
The 2th classifier AUCPR = 0.169643
The 3th classifier AUCPR = 0.000000
The 4th classifier AUCPR = 0.064474
The 5th classifier AUCPR = 0.000000
The 6th classifier AUCPR = 0.000000
The 7th classifier AUCPR = 0.000000
The 8th classifier AUCPR = 0.000000
The 9th classifier AUCPR = 0.479167
The 10th classifier AUCPR = 0.169643
precision
recall

Notice the performance for some of the classifiers not very good. This is because we only have 19 test samples (per fold). I believe if we have a larger data set, the performance between folds will be a lot better. This is because if we miss 1 or 2 the classifier will be punished harshly.

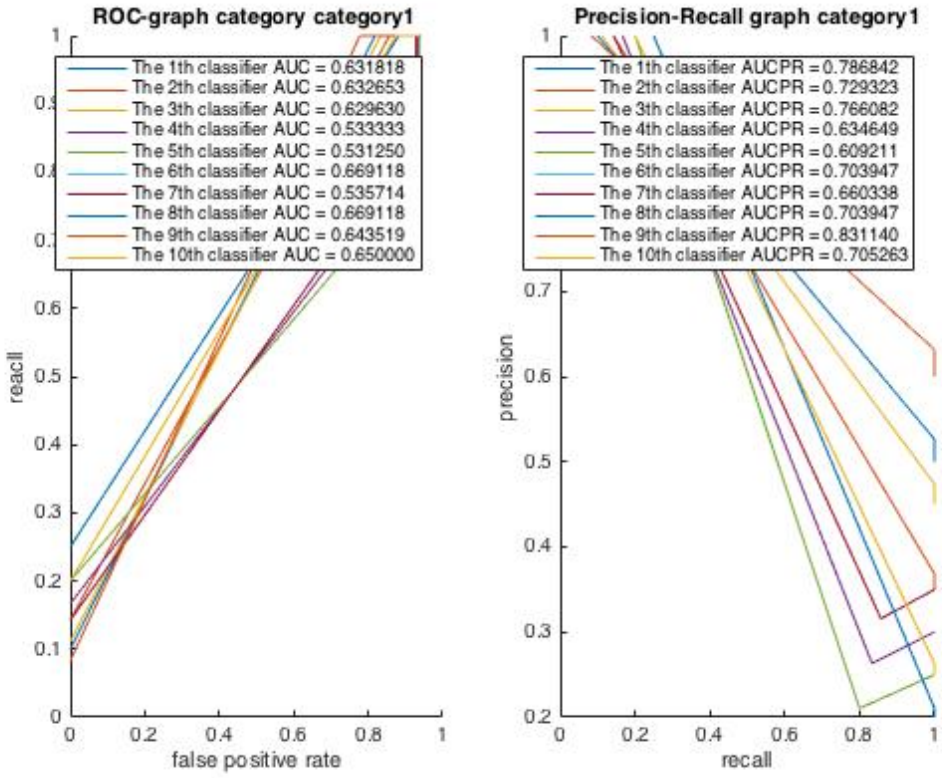
Q5
Composite:



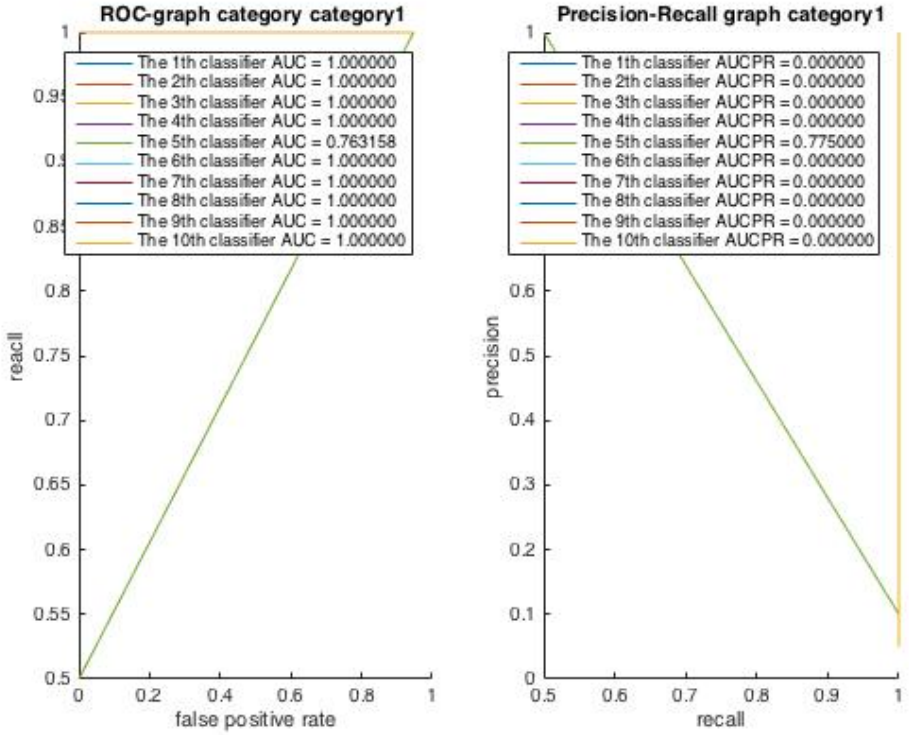

In this case, the composite one perform better than the individual ones. It's actually extremely well. 0 misclassifications. I believe this is because the composite one can better capture the feature of the data.
Therefore, we better to build a composite SVM. Plus the composite ones converge really fast, but you have more categories to classify.
NULL hypothesis is that they perform the same

Q6
Notice the training converge time for PCA is a lot longer than the composite ones as well as the individual ones. So I have to modify my code a little bit to make SMO only run 10000 times. Otherwise it is no needed to add that limit.
PCA perform worse than composite ones as well as individual ones. This is because during the process of PCA, we lose some information.

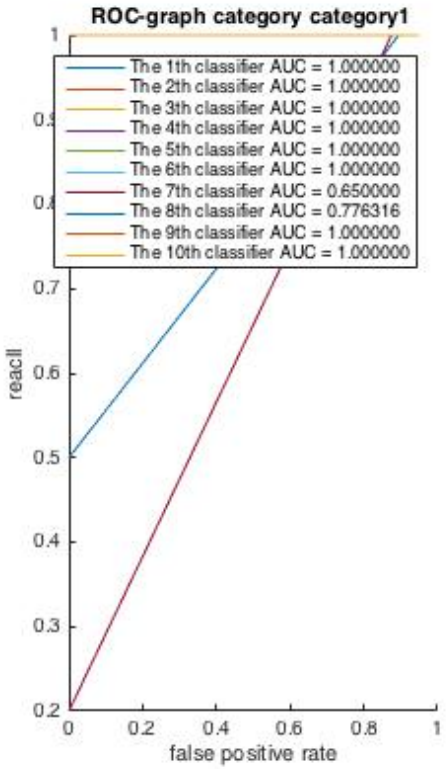
PCA medium:

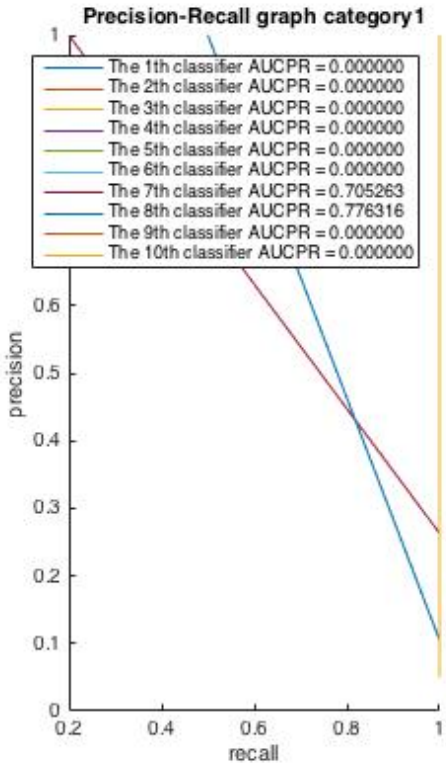



PCA strain:

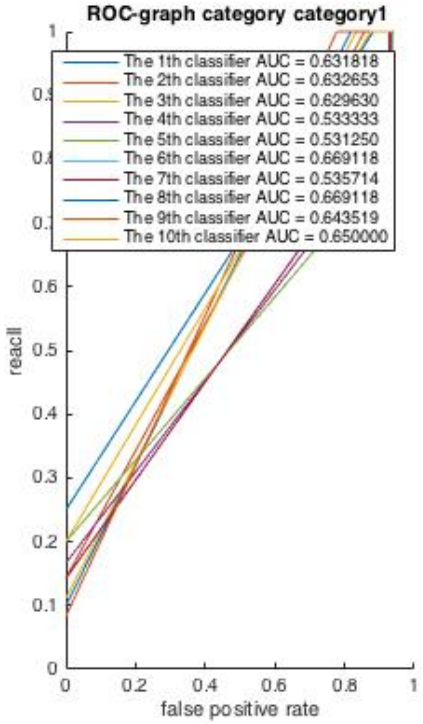


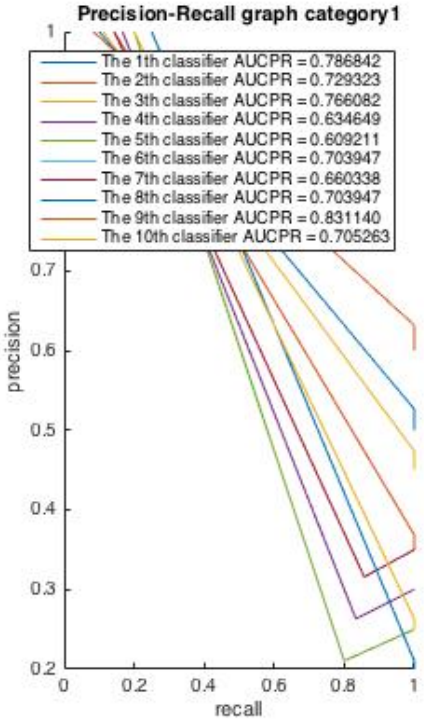

PCA GP:


ROC-graph category category1
The 1th classifier AUC = 1.000000
The 2th classifier AUC = 1.000000
The 3th classifier AUC = 1.000000
The 4th classifier AUC = 1.000000
The 5th classifier AUC = 1.000000
The 6th classifier AUC = 1.000000
The 7th classifier AUC = 0.650000
The 8th classifier AUC = 0.776316
The 9th classifier AUC = 1.000000
The 10th classifier AUC = 1.000000
reacll
false positive rate


Precision-Recall graph category1
The 1th classifier AUCPR = 0.000000
The 2th classifier AUCPR = 0.000000
The 3th classifier AUCPR = 0.000000
The 4th classifier AUCPR = 0.000000
The 5th classifier AUCPR = 0.000000
The 6th classifier AUCPR = 0.000000
The 7th classifier AUCPR = 0.705263
The 8th classifier AUCPR = 0.776316
The 9th classifier AUCPR = 0.000000
The 10th classifier AUCPR = 0.000000
precision
recall

PCA medium:


ROC-graph category category1
The 1th classifier AUC = 0.631818
The 2th classifier AUC = 0.632653
The 3th classifier AUC = 0.629630
The 4th classifier AUC = 0.533333
The 5th classifier AUC = 0.531250
The 6th classifier AUC = 0.669118
The 7th classifier AUC = 0.535714
The 8th classifier AUC = 0.669118
The 9th classifier AUC = 0.643519
The 10th classifier AUC = 0.650000
reacll
false positive rate


Precision-Recall graph category1
The 1th classifier AUCPR = 0.786842
The 2th classifier AUCPR = 0.729323
The 3th classifier AUCPR = 0.766082
The 4th classifier AUCPR = 0.634649
The 5th classifier AUCPR = 0.609211
The 6th classifier AUCPR = 0.703947
The 7th classifier AUCPR = 0.660338
The 8th classifier AUCPR = 0.703947
The 9th classifier AUCPR = 0.831140
The 10th classifier AUCPR = 0.705263
precision
recall