**Model selection**

Firstly, in the steps of process our data, we classify outputs into 2 classes,high and low(using 1 and 0 in dataset).It is a binary classification problem.Secondly, our sample size is not very big.Based on above 2 points, SVM can be a candidate model for this problem.

**Kernel selection**

SVM have 4 kind of kernel models.Based on the characteristics of our dataset, “rbf” is the most suitable model because firstly, our classification problem cannot be classified by linear model.Moreover,we only have 3 features which is relatively small and sample size is 49 students.It is not too small or not too big.

**Parameters explanation**

SVM model have a collection of parameters and after referring to online and offline resources, I decided to mainly focus on 2 parameters:C and gamma.C means penalty coefficient, if c is bigger, the model’s ability to tolerate misclassification is weaker. Gamma is a parameter that goes with kernel rbf. It determines the distribution of data when they are projected to new feature spaces.When gamma is bigger, supporting vectors are less and vice versa.

**Parameter adjustment criteria**

At the very beginning, I want to find the best parameters when accuracy score is highest.But after discussion with group members and referring to materials, I changed my criteria into ROC\_AUC score.The implicit goal of AUC is to deal with situations where you have a skewed sample distribution, and don't want to over-fit to a single class.For binary classification problem, auc is a more reliable metrics compared with accuracy score.

**Parameter adjustment strategy**

1.Initial C range list: c\_range\_list = [0.00001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000]

Initial gamma range list: gamma\_range\_list = [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000]

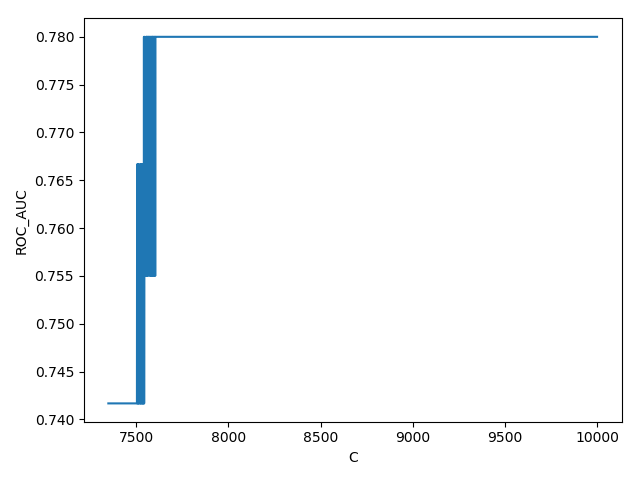
1. Find the best c, gamma in these lists using cross validation, and set cv= 5.
2. Construct a range(i.e best value-50,best value+50) based on best values in step2.
3. Observe the result of best parameter pairs(usually more than 1),if the auc score is strictly increase in current range, I will increase the range boundary to see what happen.And if the auc score is strictly go down in current range, I wll decrease the range boundary.If the current range contains the best parameter value, this value is fixed.
4. When one of C or gamma is fixed, the same as step4 ,dynamically adjust the range to retrieve the best value of another parameter.
5. Now, create a model contains above best parameters and using cross validation(cv=5) to calculate metrics:precision ,recall and accuracy scores.

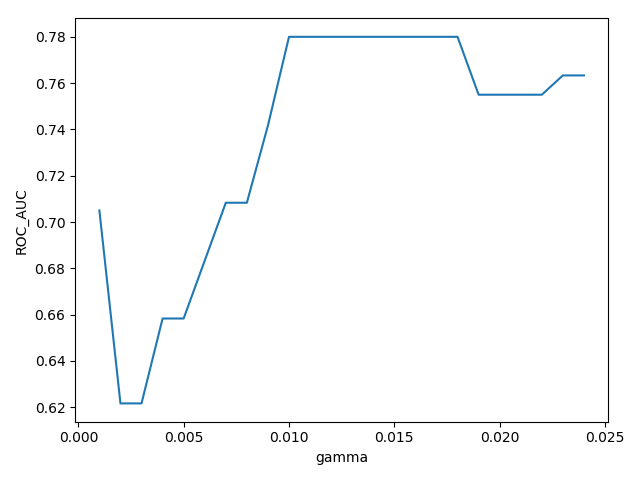
Flourishing example:

Best parameters found for different kinds of Flourishing datasets:

1.FlourishingDiffer\_Score.xlsx

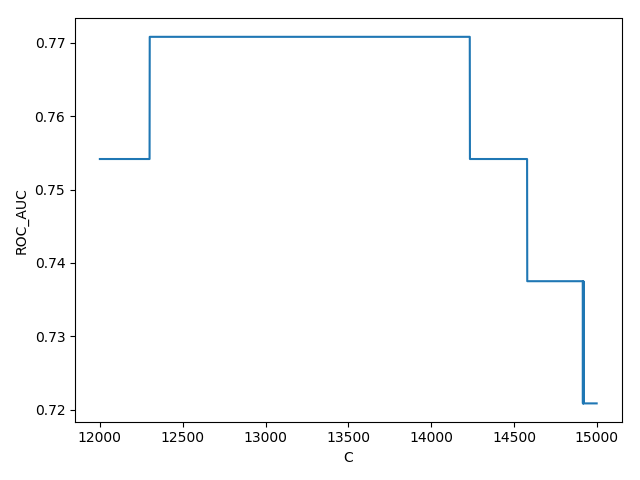
C=8000 gamma=0.01 auc = 0.78

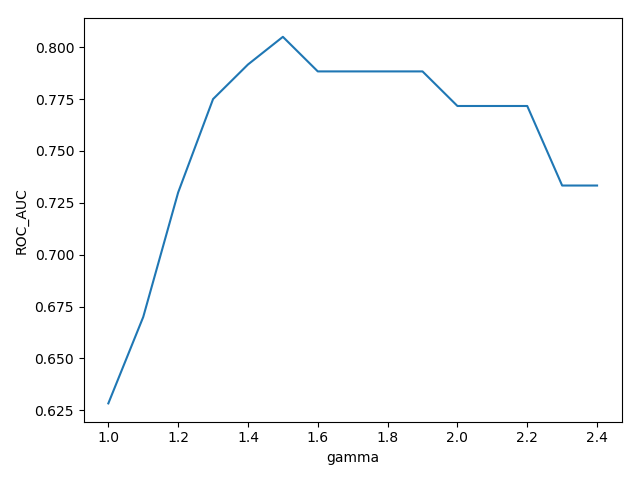




1. FlourishingDiffer\_Score\_PCA:

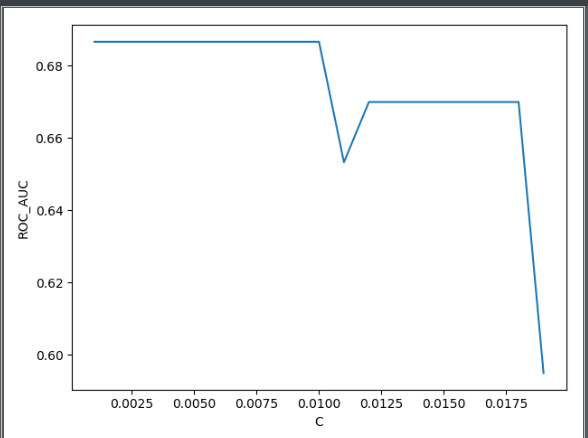
C= 14228 gamma=1.5 auc = 0.77

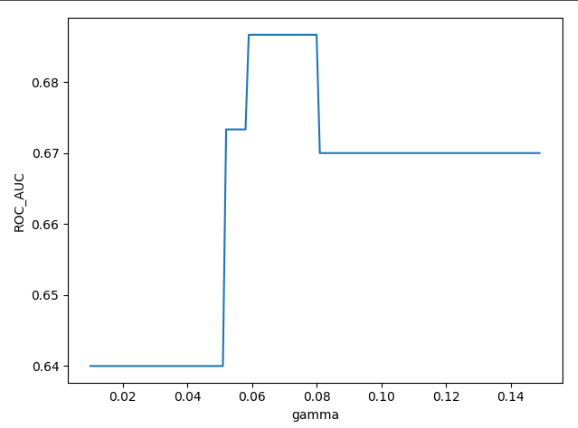




1. Simple FlourishingDiffer:

C= 0.0075 gamma=0.07 auc = 0.687





Flourishing conclusion

|  |  |
| --- | --- |
| Model SVM Kernel=’rbf’ | |
| RAW | 0.78 |
| PCA | 0.77 |
| SIMPLE | 0.687 |

From the flourishing we can see that my model performs best on raw data.So, I am going to focus on training my model on raw data.The table below shows best parameter pairs and a series of metrics for totally 9 raw datasets we generated.

Best Parameters & metrics for 9 datasets:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model SVM kernel=’rbf’ | | | | | | | |
|  | C | gamma | precision | recall | F1 | accuracy | auc |
| Flourishingpost | 8500 | 0.07 | 0.62 | 0.68 | 0.65 | 0.66 | 0.76 |
| Flourishingpre | 10500 | 0.35 | 0.64 | 0.67 | 0.65 | 0.67 | 0.76 |
| Flourishingdiffer | 8300 | 0.015 | 0.67 | 0.63 | 0.65 | 0.74 | 0.78 |
| PanasPositivepost | 0.05 | 0.78 | 0.67 | 0.62 | 0.64 | 0.65 | 0.76 |
| PanasPositivepre | 1.0 | 6166 | 0.63 | 0.62 | 0.62 | 0.64 | 0.67 |
| PanasPositivediffer | 22000 | 76 | 0.58 | 0.57 | 0.57 | 0.63 | 0.68 |
| PanasNegativepost | 40000 | 22 | 0.62 | 0.58 | 0.60 | 0.60 | 0.71 |
| PanasNegativepre | 0.0005 | 2116 | 0.55 | 0.56 | 0.55 | 0.52 | 0.59 |
| PanasNegativediffer | 36955 | 0.11 | 0.56 | 0.73 | 0.63 | 0.58 | 0.69 |

**Parameter adjustment evaluation**  
From a bunch of online and offline materials, I get to know that the range for C and gamma is significant wide,both from 10^-8 to 10^8.So, it is rather difficult or even impossible to find the global optimal parameter value.My choice is to make the order of magnitude clear first and then dynamically change the range to find out a relatively accurate one.I just adjust the parameter values within a regular range [10^-4,10^4].