HW4

0859605陳冠景

1.

1. for first canonical discriminant function: (raw canonical coefficient without

standardize)

can1 = 1.475\*(initial score)-0.314\*(midterm score)-0.557\*(final score) –

6.981

for second canonical discriminant function:

can2 = 0.155\*(initial score)-0.237\*(midterm score)-0.054\*(final score)-

11.409

\*Use the CDA$scores function in candisc to calculate the score in second

canonical function score for each case.

\*

|  |  |
| --- | --- |
| Table 1. | |
| PC\_Ability | Can2 |
| good | 0.609028 |
| excellent | -1.57256 |
| good | 2.159603 |
| excellent | -0.24087 |
| bad | 0.879041 |
| good | 0.405576 |
| bad | -0.52163 |
| good | 0.606611 |
| bad | -0.03394 |
| excellent | 0.371087 |
| good | -1.26269 |
| bad | -1.33972 |
| excellent | -0.82855 |
| good | 0.06363 |
| excellent | 0.705388 |

1. combined group plot (figure 1.)

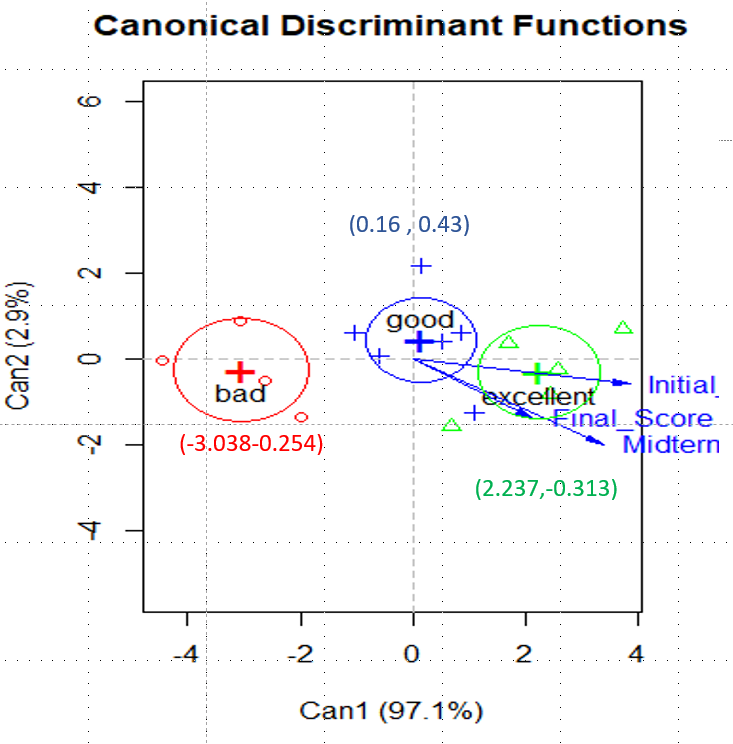


Figure 1. group plot from canonical discriminant functions.

As the result shows, the group of bad PC ability has negative first canonical variate score , group of good PC ability has the highest second canonical variate score among three group ,and group of excellent PC ability has the lowest second canonical variate score among three group.

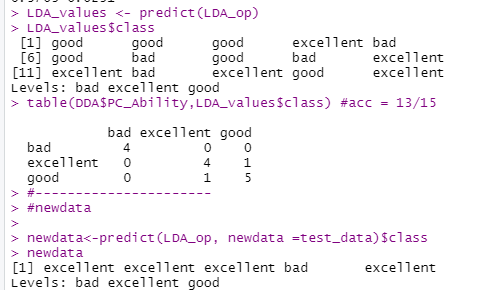
1. Fisher’s linear discriminant functions in equations.
2. fisherbad = -0.964\*(initial score)+2.454\*(midterm score)+0.54\*(final score)-77.759
3. fisherexcellent = 0.146\*(initial score)+2.187\*(midterm score)+0.272\*(final score)-113.946
4. fishergood = -0.179\*(initial score)+2.122\*(midterm score)+0.338\*(final score)-87.337

\*calculate for each case

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2. | | | |
| Sample PC\_Ability | bad | excellent | good |
| 1 | 73.404 | 70.200 | 75.442 |
| 2 | 105.893 | 111.889 | 111.938 |
| 3 | 61.040 | 63.959 | 67.907 |
| 4 | 98.411 | 114.438 | 111.498 |
| 5 | 61.865 | 47.935 | 57.593 |
| 6 | 79.697 | 84.768 | 86.606 |
| 7 | 80.038 | 68.487 | 76.200 |
| 8 | 78.849 | 85.604 | 86.924 |
| 9 | 66.339 | 45.134 | 56.998 |
| 10 | 85.043 | 96.318 | 95.689 |
| 11 | 103.965 | 112.181 | 111.580 |
| 12 | 92.221 | 84.121 | 89.886 |
| 13 | 102.367 | 117.707 | 114.614 |
| 14 | 81.527 | 80.622 | 84.566 |
| 15 | 91.315 | 113.390 | 108.751 |

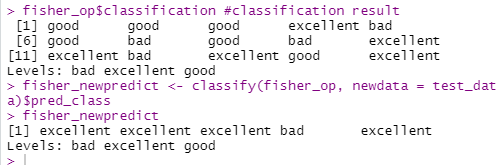
1. Fit new data into linear discriminant analysis function and fisher’s linear discriminant function to predict their PC ability.

\* Linear discriminant analysis



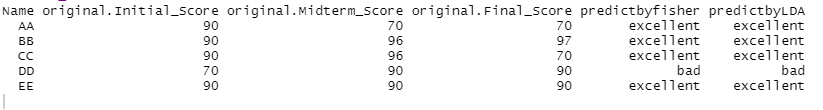
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3. Result of new data from Linear discriminant analysis | | | | |
| Name | Initial\_Score | Midterm\_Score | Final\_Score | Predict PC\_ability |
| AA | 90 | 70 | 70 | excellent |
| BB | 90 | 96 | 97 | excellent |
| CC | 90 | 96 | 70 | excellent |
| DD | 70 | 90 | 90 | bad |
| EE | 90 | 90 | 90 | excellent |

\* Fisher’s linear discriminant functions

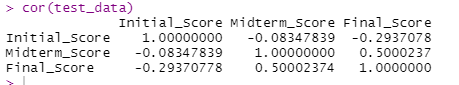


We get same result as table 3 above. That only student DD is in the bad PC ability category, the others are in the PC excellent ability category.

The predicted memberships for the new cases from linear discriminant analysis function and fisher’s linear discriminant analysis are totally consistent.

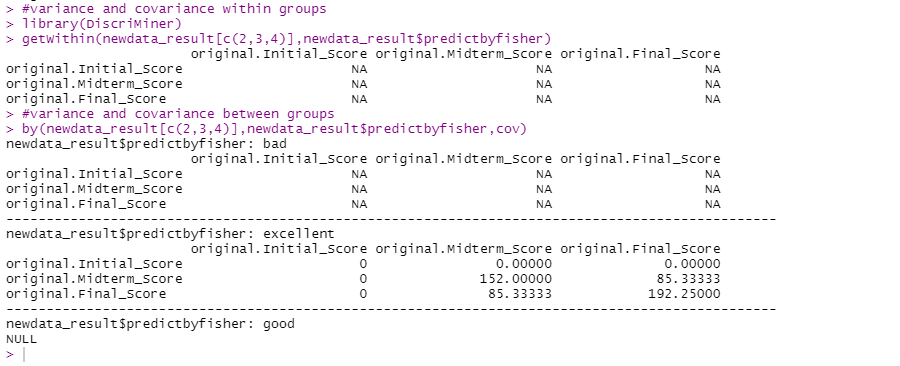


1. The relationship of three exam scores from the correlation analysis.

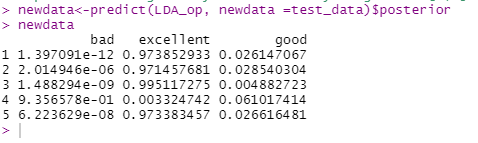


The initial score is negatively correlated with midterm score and final score, but the midterm score is positively correlated with final score.

As a result , I try to do the variance and the covariance within and between the group, but there is only four excellent ability samples and one bad ability sample. Since only a few number of new data , the result shows null result.



By linear discriminant analysis, use *posterior* to show the function scores in three categories.



Since only sample 4 has higher good ability score than excellent ability score, the other 4 samples all have higher excellent ability scores than the other 2 abilities score. As a result , sample AA,BB,CC,EE are more likely in excellent PC ability group, and sample DD is more likely in good PC ability group. Otherwise, by comparing their original scores, except DD, the others who have same score in initial score are all in excellent group. Sample DD who has lower score in initial exam is in the good group, not in the excellent group as others sample.

2. use the HW4\_data to do RF,SVM,ANN

First , split data to train and test data set according to 70% of raw data in train set and 30% of raw data in test set.

Random Forest model(RF):

confusion matric:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 4. confusion matrix of RF result in train data | | | | |
| Pre refer | bad | excellent | good | class.error |
| bad | 1 | 0 | 2 | 0.666 |
| excellent | 0 | 1 | 2 | 0.666 |
| good | 1 | 1 | 3 | 0.4 |

By *importance,* table 5 shows that initial score is most important in bad and good group and the midterm score is most important in excellent group.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 5.Importance about 3 scores in 3 category | | | |
|  | bad | excellent | good |
| Initial\_Score | 9.201248 | 2.144779 | 3.217012 |
| Midterm\_Score | 0.254473 | 7.904809 | -3.40824 |
| Final\_Score | -1.09676 | -1.42985 | -5.10897 |

confusion matric:

|  |  |  |  |
| --- | --- | --- | --- |
| Table 6. confusion matrix of RF result in test data | | | |
| Pre refer | bad | excellent | good |
| bad | 1 | 0 | 0 |
| excellent | 0 | 1 | 0 |
| good | 0 | 1 | 1 |

Model evaluation : accuracy score = 0.75

Other index:

|  |  |  |  |
| --- | --- | --- | --- |
| Table 7. evaluation index of RF | | | |
| Class: | bad | excellent | good |
| Precision | 1 | 1 | 0.5 |
| Recall | 1 | 0.5 | 1 |
| F1 | 1 | 0.6667 | 0.6667 |

In bad PC ability group, the prediction is highly correct, so the precision index ,F1and recall index are all high. In the excellent PC ability group , since one sample is incorrectly predict ,the recall index is only 0.5.

Support vector machine(SVM):

|  |  |  |  |
| --- | --- | --- | --- |
| Table 8. confusion matrix of SVM result in train data | | | |
| Pre refer | bad | excellent | good |
| bad | 1 | 0 | 0 |
| excellent | 0 | 1 | 0 |
| good | 0 | 1 | 1 |

In SVM, 10-fold cross-validation is used, the accuracy is 63.636.

Test data by SVM ,the confusion matrix show in table 9.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 9. confusion matrix of SVM result in test data | | | |
| Pre refer | bad | excellent | good |
| bad | 1 | 0 | 0 |
| excellent | 0 | 1 | 0 |
| good | 0 | 1 | 1 |

Model evaluation : accuracy score = 0.75

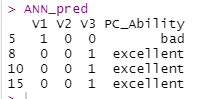
Other index:

|  |  |  |  |
| --- | --- | --- | --- |
| Table 10. evaluation index of SVM | | | |
| Class: | bad | excellent | good |
| Precision | 1 | 1 | 0.5 |
| Recall | 1 | 0.5 | 1 |
| F1 | 1 | 0.6667 | 0.6667 |

The precision, recall and F1 evaluation indices in support vector machine model are in accordance with those in Random forest model.

Artificial neural network (ANN):

Test data in ANN prediction.



The confusion matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Table 11. confusion matrix of ANN result in test data | | | |
| Pre refer | bad | excellent | good |
| bad | 1 | 0 | 0 |
| excellent | 0 | 2 | 1 |
| good | 0 | 0 | 0 |

Model evaluation : accuracy score = 0.75

Other index :

|  |  |  |  |
| --- | --- | --- | --- |
| Table 10. evaluation index of ANN | | | |
| Class: | bad | excellent | good |
| Precision | 1 | 0.6667 | NA |
| Recall | 1 | 1 | 0 |
| F1 | 1 | 0.8 | NA |

The bad group has better prediction index as compare to the other two group as the RF and SVM model.

The neural network has 2 hidden layers, and the error is 3.63, which is quite large.(fig 2.)

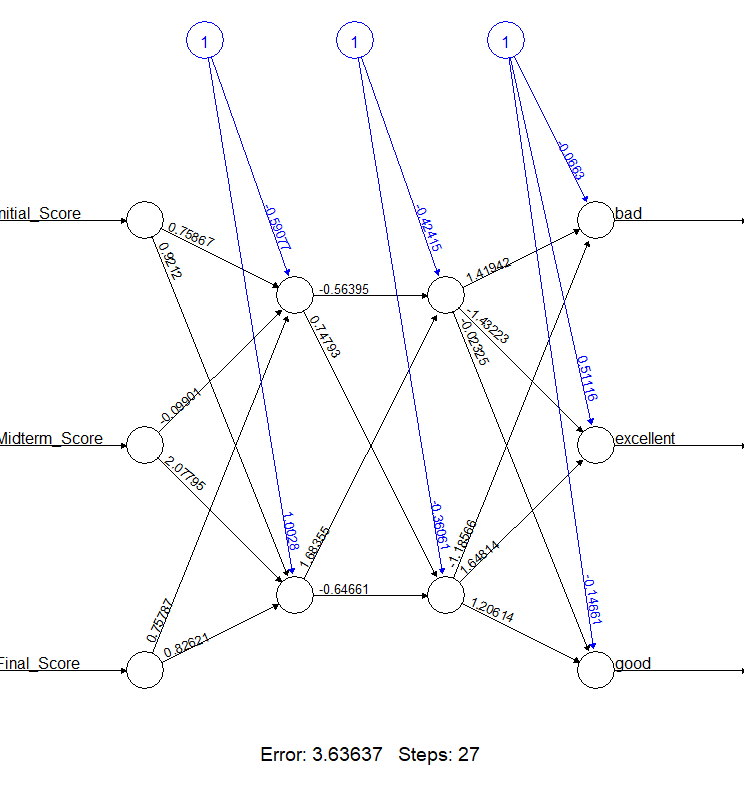


Figure 2.ANN

In three supervised machine learning model , the accuracy of test data are all 0.75, but the other evaluation indices are not completely the same. The good ability group in ANN has poor prediction than the other two classification models.

In excellent ability group, ANN model has higher F1 scores(F1 =0.8) than SVM and RF models(F1 = 0.667). We choose F1 to make a comparison between different machine learning models since it includes precision and recall index at the same time.

Since ANN do not predict the good ability group , the SVM and RF models have same prediction indices in 3 PC ability group, it is hard to choose one classifier model between these three classifiers to represent as the classification model in th