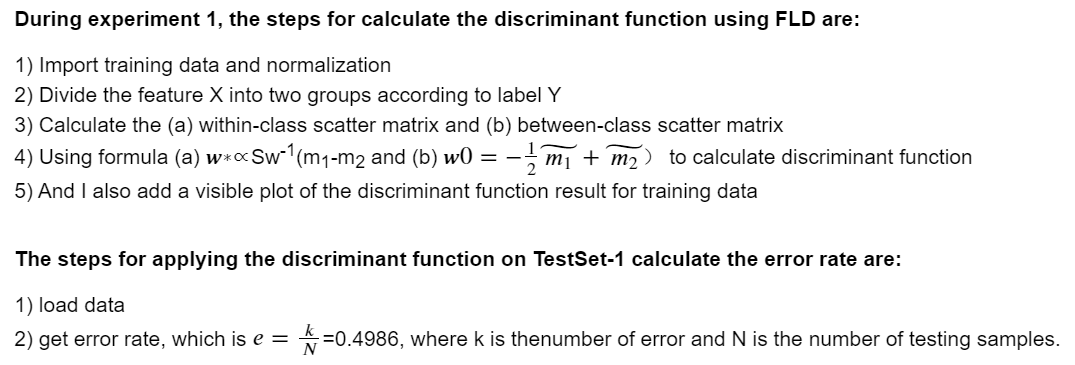
# Computer Exercise 2

2022312873李佳璇

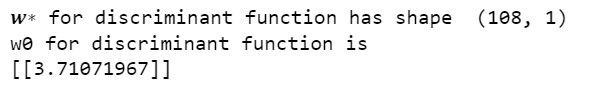
**Notation: The content in word it totally the same as .ipynb document running result.**

## Experiment 1

**1) Experiment Steps:**



**2) Discriminant function**



**3) Classification**

On training set:

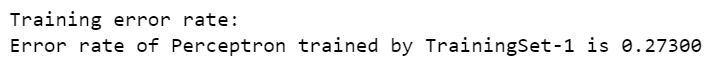


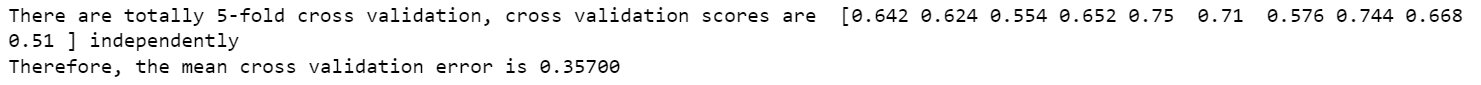
On testing data:

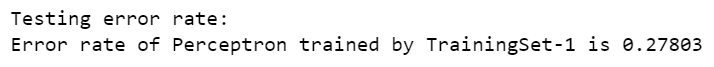
Error rate of FLD discriminant function is **0.49863**

## Experiment 2

1) **The training error rate is 0.27300, cross validation error is 0.35700, and testing error rate is 0.27803**

****

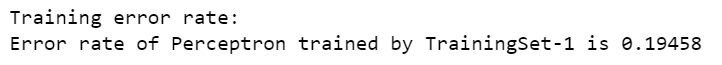
****

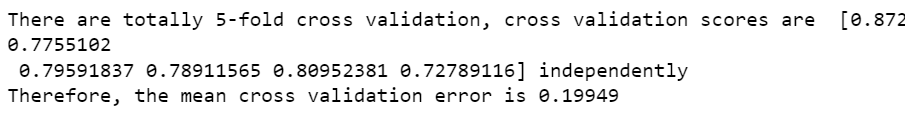
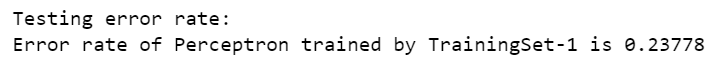


We can see that the error rates in training set and testing set are almost the same (training is slightly lower than testing), which means the classifier has a good generalization performance.

And cross validation error is and approximation of testing error rate based on training data, but here maybe because when divided the training data into 10 groups, the amount training data decrease, so the value is higher than the other two.

2) **The training error rate is 0.19458, cross validation error is 0.19949, and testing error rate is 0.23778**



We can see that the error rates in training set and cross validation error are almost the same (training is slightly lower than cross validation), which means the classifier has a good generalization performance see from the 0.005 difference.

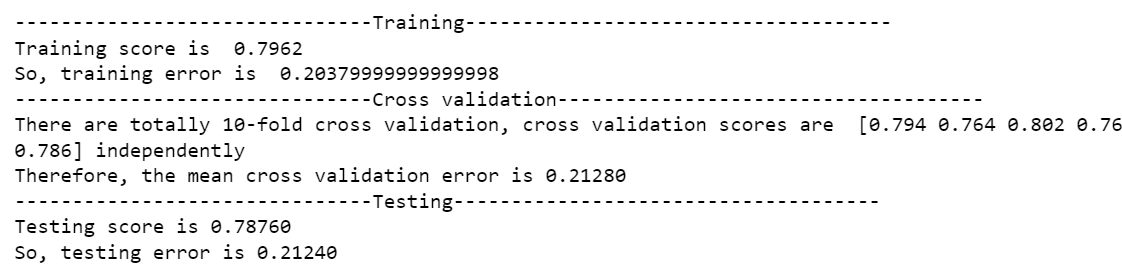
But the testing error is higher than the other two by approximately 0.038, which means the classifier is still a little bit weak on the data that haven't seen before.

3) The result shows that the error rates onTrainingSet-2 are smaller than that on TrainingSet-1.

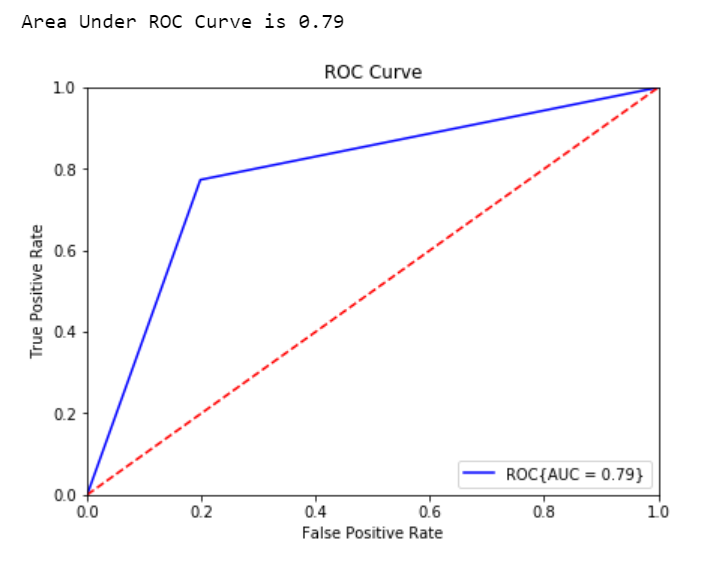
As Perceptron shows different error rate on different datasets, it proves that TrainingSet-2 is better than TrainingSet-1, when doing the classification of 'hospital\_death'. For training, TrainingSet-2 performs better.

## Experiment 3

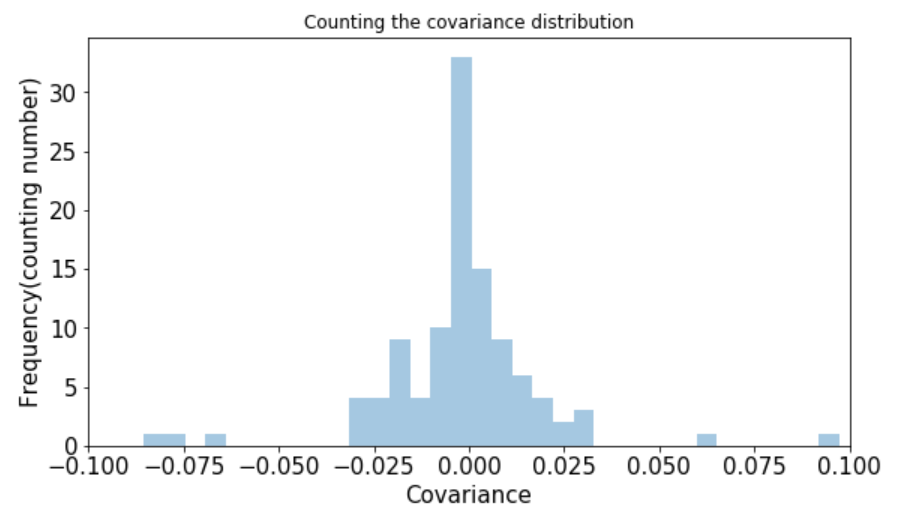
1) **training error is 0.20380, mean cross validation error is 0.21280, and testing error is 0.21240**



2) ROC curve shows below



3) Influence distribution shows below:



From the plot above, we can see that most of the features only have small correlation with patients' survival.

But there are still some features with significant influence

-------------------------------significant influence-------------------------------------

There are 2 features has high **positive correlation** with final result:

***['intubated\_apache', 'ventilated\_apache']***

There are 3 features has high **negative correlation** with final result:

***['gcs\_eyes\_apache', 'gcs\_motor\_apache', 'gcs\_verbal\_apache']***