MSBD6000B Project 1

Data Preprocessing

For this dataset , there are large amounts of zeros in there, the feature is really sparse , for some specific feature even almost all rows is zero, so I think it is possible to get a better performance when we shrink the dimension of features, so I tried the some feature selection method, including the PCI and Chi-square test, variance in feature and selection from models.

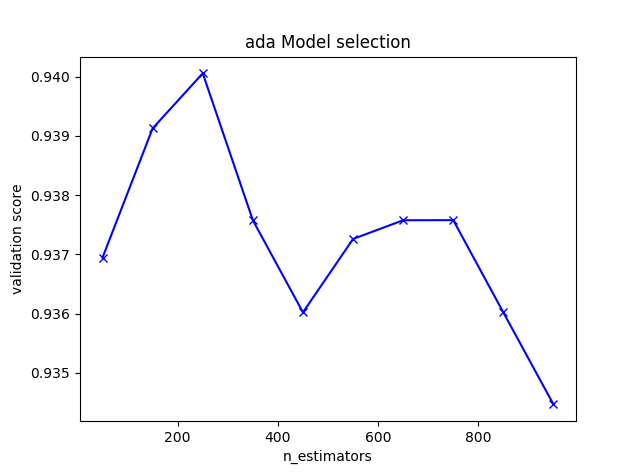
Finally I choose to sure variance and selection the best feature using Logistic regression with L1 Penalty.

Experiment Process:

## Adaboost classifier

I tried different estimators number , it turns out the best estimators number when base classifier is decision tree is

The best validation score the Adaboost classifier(weak classifier = decision tree) is 0.941 when we weak estimator number = 200



## SVM classifier

For Support vector machine I tried to use grid-search to find out the best gamma and penalty parameters, so I search the 2 parameter in ,

So according to my experiment as we have **Gaussian RBF kernel** the best parameters are:

{'C': 100000, 'gamma': 1e-06} with a validation score of 0.94

the best score Support vector machine can get is around 0.94

## Fully connected neutral network classifier

I implemented an 5 layers fully connected neural networks by Keras, After experiments, it is seems that when use the neutral network classifier feature selection is not necessary (feature selection may lead to a worse result).

**Do the cross validation the tuning the parameter of networks:**

Firstly I tried five layers NN:

For [64,128,128,64,64] six layers structure epoch number =500 :

|  |  |  |  |
| --- | --- | --- | --- |
| Drop-out | 0 | 0.3 | 0.6 |
| Valid score | 0. 93581781 | 0. 94824016 | 0.93892339 |

For [64,128,128,128,64] six layers structure epoch number =1000 :

|  |  |  |  |
| --- | --- | --- | --- |
| Drop-out | 0.3 | 0.4 | 0.6 |
| Valid score | 0.94927536 | 0. 95031055 | 0. 93064183 |

For [64,128,128,64,64] five layers neutral network, when not use dropout , we can get following learning curve with maximum 0.9919 acc, but only 0.9358 validation score, which may means model is over-fitted, so I tried 0.3 and 0.6 dropout, I turns out 0.3 dropout can get an better validation score.

Then I also tried out some different layer numbers an picked out the best parameter:

[64,128,128,128,64] , Drop-out = 0.4

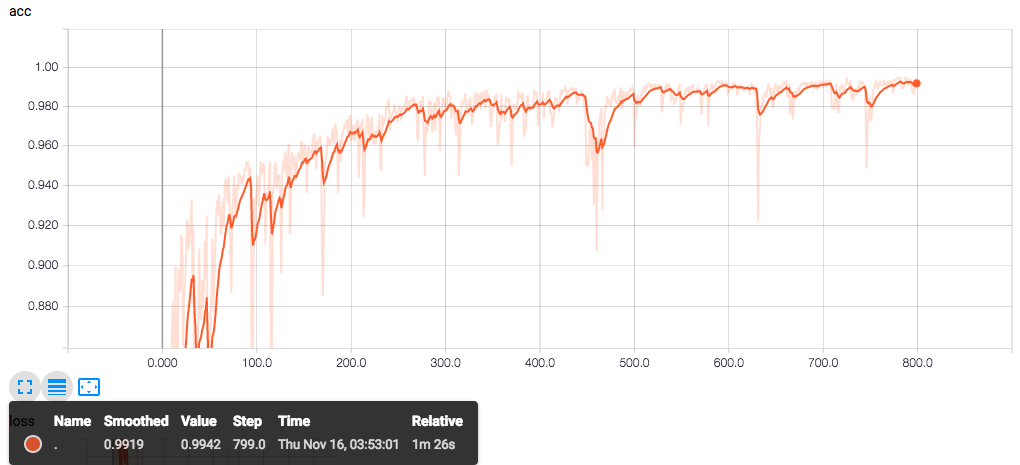
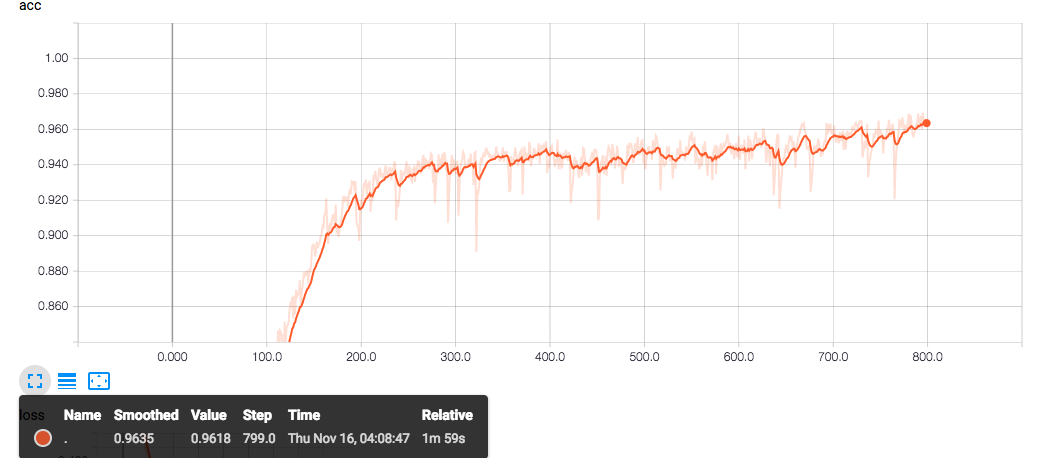
Figure1: [64,128,128,64,64] , dropout = 0, epoch number =800 (over fitted)

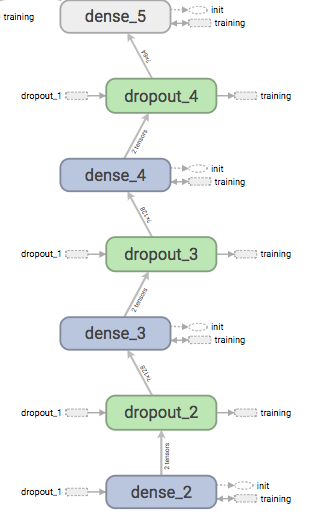
Figure2:[64,128,128,128,64] , dropout = 0.3, epoch number =800 (Under fitted)



So our work is:

( 1 ) Define the 6 layer fully connected neutral network.

( 2 ) search the best dropout percentage(bias and variances balance)



So after I tuned the parameters for these 3 algorithms , it turns out the validation score is quite close , and finally I decide to pick the best one: **Fully connected neutral network classifier.**