

## HW-5 : Random Forest

組別27

0310515葉尚昀

0310527李韋辰

0310519陳柏諺

降維：

Pca：

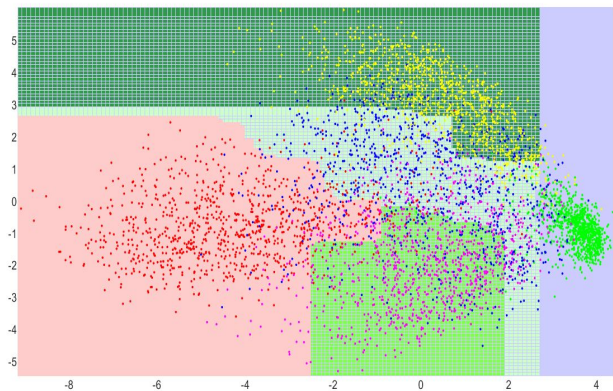

下圖為PCA 前十個主成分所佔的比例，由此可見，其所含的比例皆很少，Lose 的data 太多，於是我們認為這次的題目不適合用PCA來降維

```
[ 0.14783296  0.08168733  0.06309249  0.059925    0.05106545  0.03875519
 0.03379848  0.02658789  0.02549452  0.02191384]
```

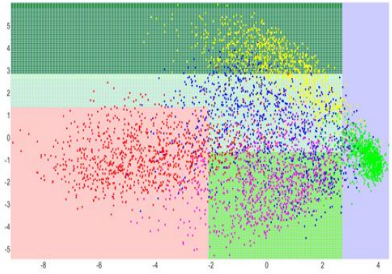
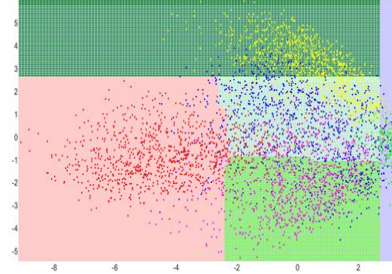
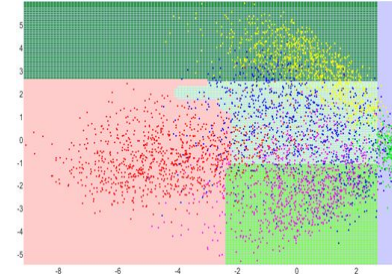
但是想藉由pca 降維經過Random Forest所繪出的decision region 來大致探討The number of decision trees ,The minimum number of samples per leaf node 和The fraction of samples所造成的影響。

### □pca vs neural network

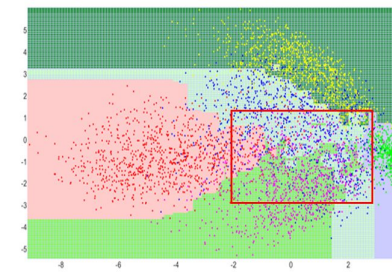
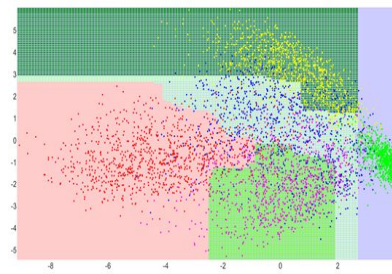
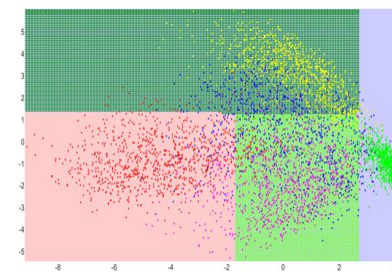
藉由pca降成2維代入andom forest 繪製出decision region 和利用neural network 取中間hidden layer (node = 2 ) 當成training data 代入random forest 所繪製而成的decision region 進行比較。

The minimum number of samples per leaf node=100 The fraction of samples=1 /The number of decision tree=100		
降維	pca	neural network
decision region		
error rate	551/2500	533/2500
討論	我認為兩者降成二維呈現的效果都不佳，因為當將原本的784維度降成2維度，損失過多的特徵值，很難以2維的資訊較完整表現原有的資料。	

## □ The number of decision trees

固定 The minimum number of samples per leaf node 和 The fraction of samples. <b>改變 The number of decision trees</b> 來探討其對於結果的影響。 The minimum number of samples per leaf node=500/The fraction of samples=1			
The number of decision trees	1	50	1000
decision region(pca)			
error rate (pca)	661/2500	636/2500	633/2500
討論	當 The number of decision trees=1 時，因為樹的數目過少，而只由過少的樹來決定其 decision region，因此易受到效果較差的樹影響，所以其 decision region 分出來的效果較差，error rate 較大；反之，當 The number of decision trees 增加時，利用多顆樹來比重來決定其 decision region，不易受到效果較差的樹影響，所以效果較佳，error rate 較低。		

## □ The minimum number of samples per leaf node

固定 The number of decision trees 和 The fraction of samples. <b>改變 The minimum number of samples per leaf node</b> 來探討其對於結果的影響。 The number of decision trees=50/The fraction of samples=1			
The minimum number of samples per leaf node	10	100	1000
decision region(pca)			
error rate (pca)	553/2500	550/2500	736/2500
討論	當 The minimum number of samples per leaf node 過小 ( <b>The minimum number of samples per leaf node=10</b> ) 時，我認為發生 overfitting 的問題或者是說可以區分出離群點 (outlier)，如同上方左圖紅色方框，明顯可以看到有一小區塊的紅色和綠色的 decision region；但當 the minimum number of samples per leaf node 過大時，便無法區分出完整的 decision region，因為當切出來的數目達到設定的值時，就不會在繼續進行切割而導致無法表現出其真正的 decision region。		

## □ The fraction of samples

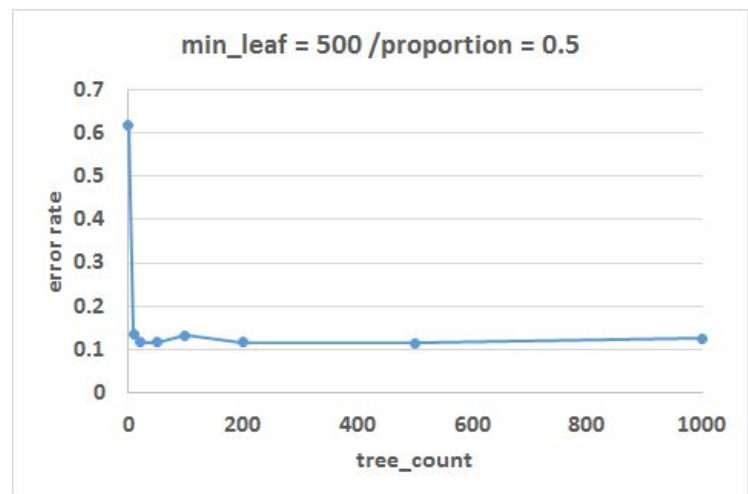
表格中的decision tree 為 利用neural network 取中間hidden layer (node = 5 ) 當成training data 代入random forest 所繪製而成的decision tree , 同時也去觀察The fraction of samples 的變動對於error rate 的影響。

固定 The number of decision trees和The minimum number of samples per leaf node. 改變The fraction of samples 來探討其對於結果的影響。 The number of decision trees=100 / The minimum number of samples per leaf node=1000			
The fraction of samples	0.4	0.5	0.7
Decision Tree			
	Error rate: 0.608	Error rate: 0.6064	Error rate: 0.44
	Error rate: 0.626	Error rate: 0.6116	Error rate: 0.4328
	Error rate: 0.6248	Error rate: 0.6052	Error rate: 0.44
error rate	460/2500	625/2500	347/2500

## Parameter adjustment

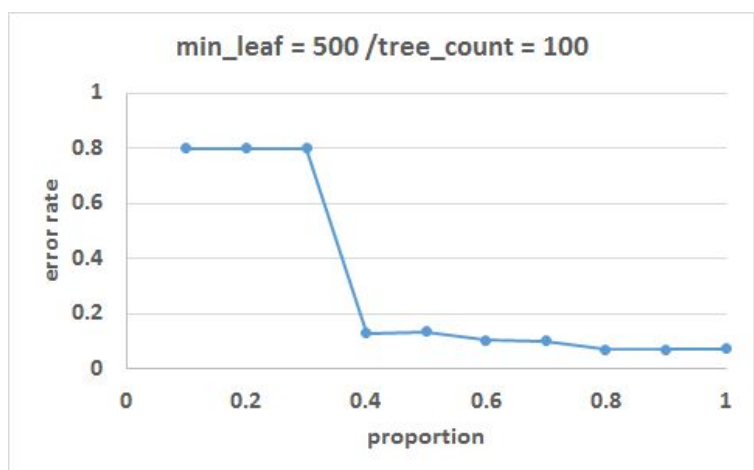
利用neural network 取中間hidden layer (node = 5 ) 當成training data 代入random forest 去觀察不同參數的數值對於error rate 的影響。

min leaf node= 500  
the fraction of sample = 0.5  
以決定樹的數量與誤差百分比作圖



當樹木的數量越少時誤差越大，由於單一樹木的判斷不佳，故當樹木數量達到某個臨界值之後，誤差即會有明顯下降，而在樹木有足夠數量之後，誤差便較沒有變動的趨勢，此時增加樹木僅增加運算的時間而沒有明顯的準確度上的提升

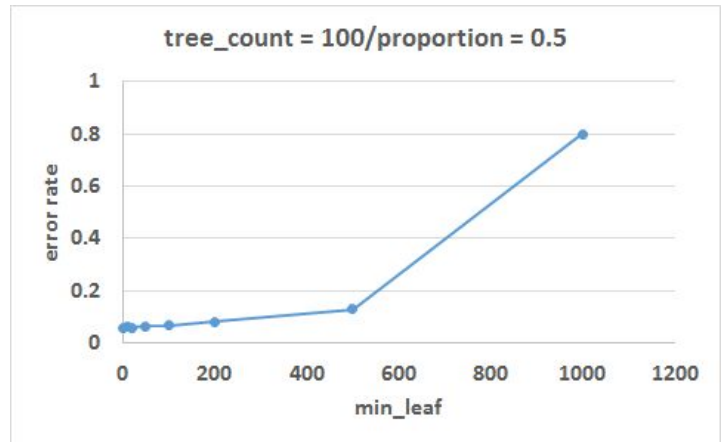
min leaf node = 500  
tree = 100  
以bagging proportion與誤差率作圖



當bagging proportion達到0.4以上之後，誤差率有明顯下降的趨勢，推測為當proportion過低時，使得每次取到的資料關聯度下降，因此在Random forest的判斷時才會誤差過大。



tree = 100  
bagging proportion = 0.5  
以決定樹的葉子數量與誤差率作圖



由上圖可知，隨著葉子數量的增加，誤差率有隨之上升的趨勢，此現象也很直覺。葉子數量的降低相當於每棵樹的分類也都更加準確，因此整體的誤差也會跟著下降。