G53MLE Assignment 3: Dimensionality Reduction

Date: 2nd Dec 2016

Group 3 Members:
Zisi Ye
Yiming Zhang
Zhihan Zhang
Afraz Hussain
Zher Shin Tan

1. Introduction

In the assignment 3, in order to achieve dimensionality reduction, the group implemented two methods correlation based feature selection and principal component analysis to the facial point data. Applying these two methods both result in a smaller features(dimensionality), but they achieve it by two different ways. The CFS method calculates all the features in the input data using CFS formula and select certain features according to their perforamce of CFS value. However, the PCA method converts a set of feature point data to a set of new values(principal compoments). To compare these two methods, the group applied the results of two methods to the decision trees which the group had built in assignment 2. The result of confusion matrix, precision rate, recall rate and F_1 measure rate will illustrate which method has a better performance.

2. Run the code

'Correlation.m': this function file calculates Pearson's (product-moment) correlation coefficient between two objects.
'CFS.m': this function file implements the correlation based feature method.
'cw3_cfs.m': this file implements the 10-fold cross validation and decision tree while the input data has been processed by the correlation based feature method. Run this file, the 'result' in the workspace is the confusion matrix for CFS method. It may take a few minutes to run once.
'cw3_pca.m': this file implements the 10-fold crooss validation and decision tree while the input data has been processed by the principal component analysis method. Run this file, the 'result' in the workspace is the confusion matrix for PCA method. It may take a few minutes to run once.

'evaluation.m': The result of recall, precison and F1 measure value are all in the workspace after running it.
(Note: the group also submitted two files named 'cw3_cfs1' and 'cw3_pca1' which looks more concise in coding, but they are actually the same as 'cw3_cfs' and 'cw3_pca'. However, 'cw3_cfs1' and 'cw3_pca1' are **not recommended** to run because they will cost 10- 15 minutes to run once. Please run 'cw3_cfs' and 'cw3_pca' for testing purpose.)

3. Results

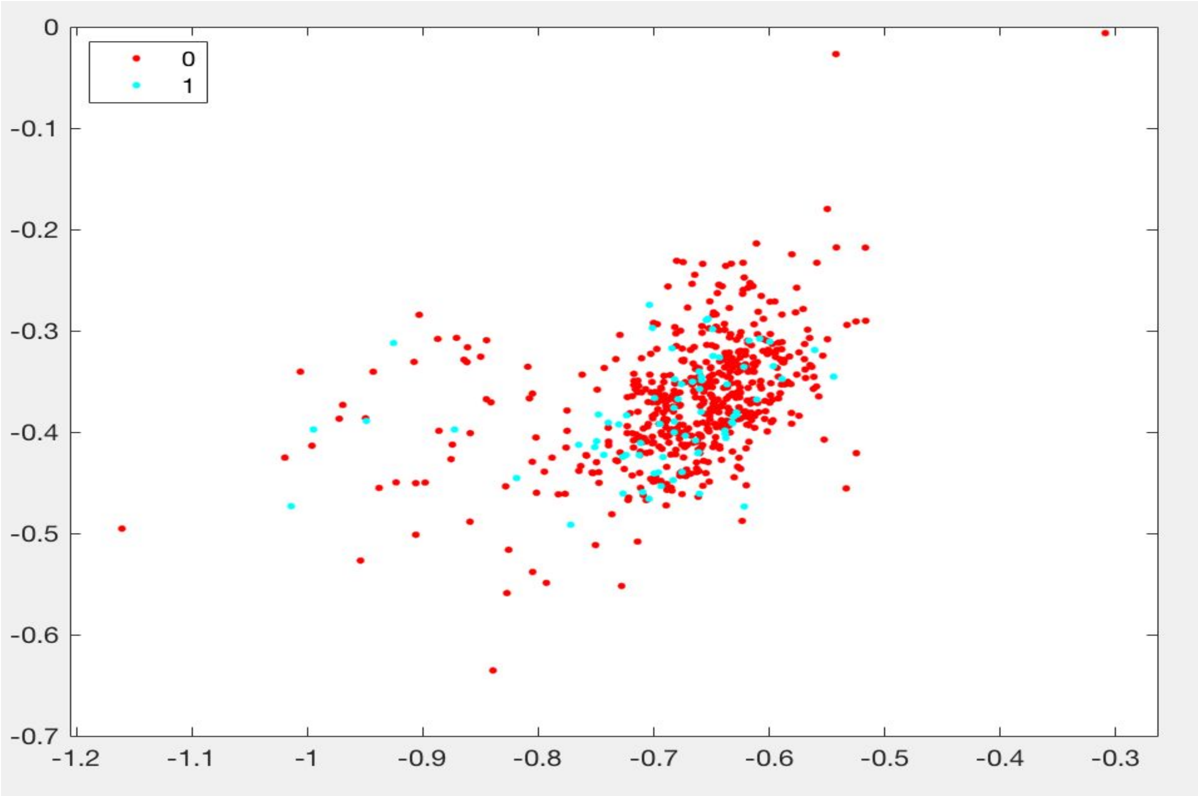
3.1 Correlation Based Feature Selection

The CFS in report is implemented in a 1-vs-all manner. That is, in this case, applying CFS for each emotion. Target classes are therefore become binary: 1 for the specific emotion to be recognized, 0 for other emotion. To decide on which feature to be selected, our code evaluate the Pearson's correlation coefficient between the target classes and newly added feature. The correlation coefficients between any pair of selected features are also calculated. The purpose of these calculations is to find a feature that maximize the correlation between classes and features and minimize the correlation between selected features. Our algorithm keeps selecting best features by proposed criterion until defined stop condition is reached.

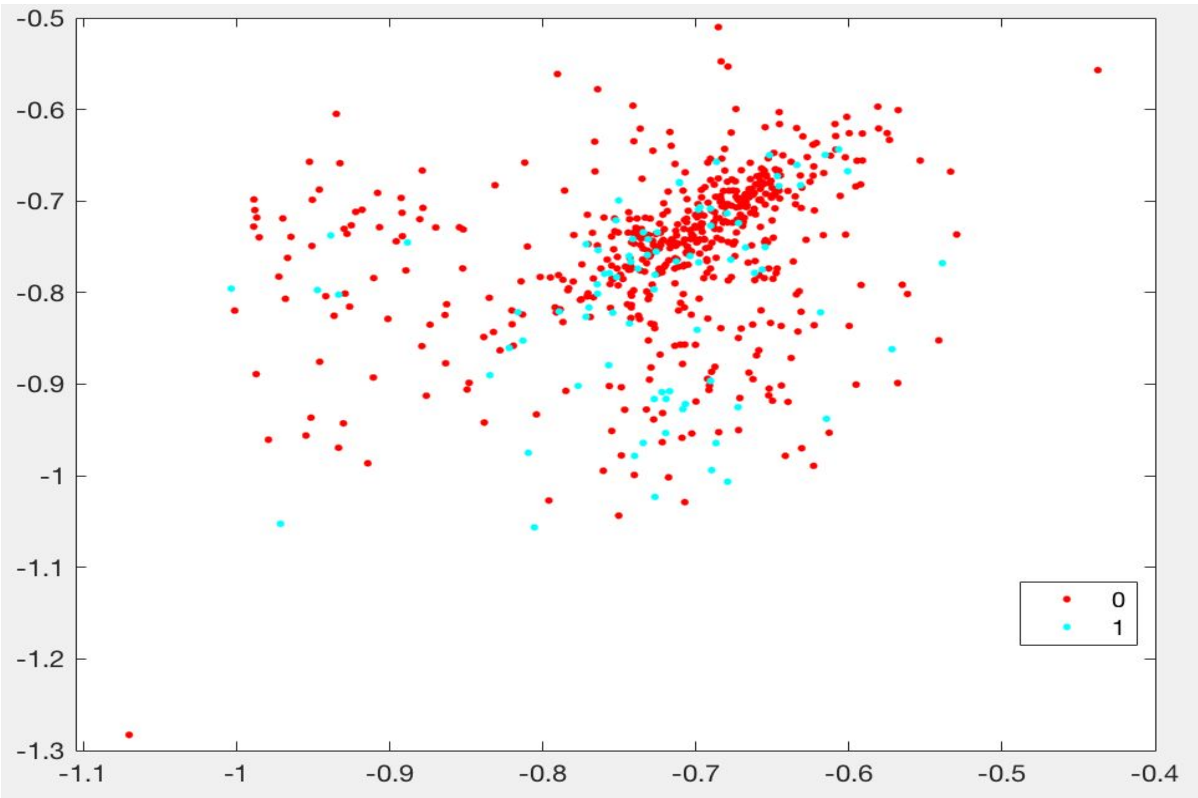
As the value of CFS is the prime criterion in feature selection, it is sensible to stop selecting new features when the value of CFS starts to converge or decrease. Hence our algorithm will stop the selection if a number of continous decreases is observed, and the features before n continous decreases are selected. (features that decreases CFS are excluded) After tested with few possible number, 3 is chosen to be the maximum of continous decrease. In our experiment, setting the maximum to 3 appears to give a good approximation of global maximum of CFS function.

The most important features are the features that have largest improvement on the value of CFS, considering the main purpose of CFS is to maximize the correlations between classes and atttributes while minimize the correlations among chosen attributes. Hence, the most important feature for emotion anger(label 1) is 4th feature and it has selected 61 features, the most important feature for emotion disgust(label 2) is 3rd feature and it has selected 70 features, the most important feature for emotion fear(label 3) is 3rd feature and it has selected 40 features, the most important feature for emotion happiness(label 4) is 6th feature and it has selected 59 features, the most important feature for emotion sadness(label 5) is 2nd feature and it has selected 52 features, and the most important feature for emotion surprise(label 6) is 2nd feature and it has selected 49 features.

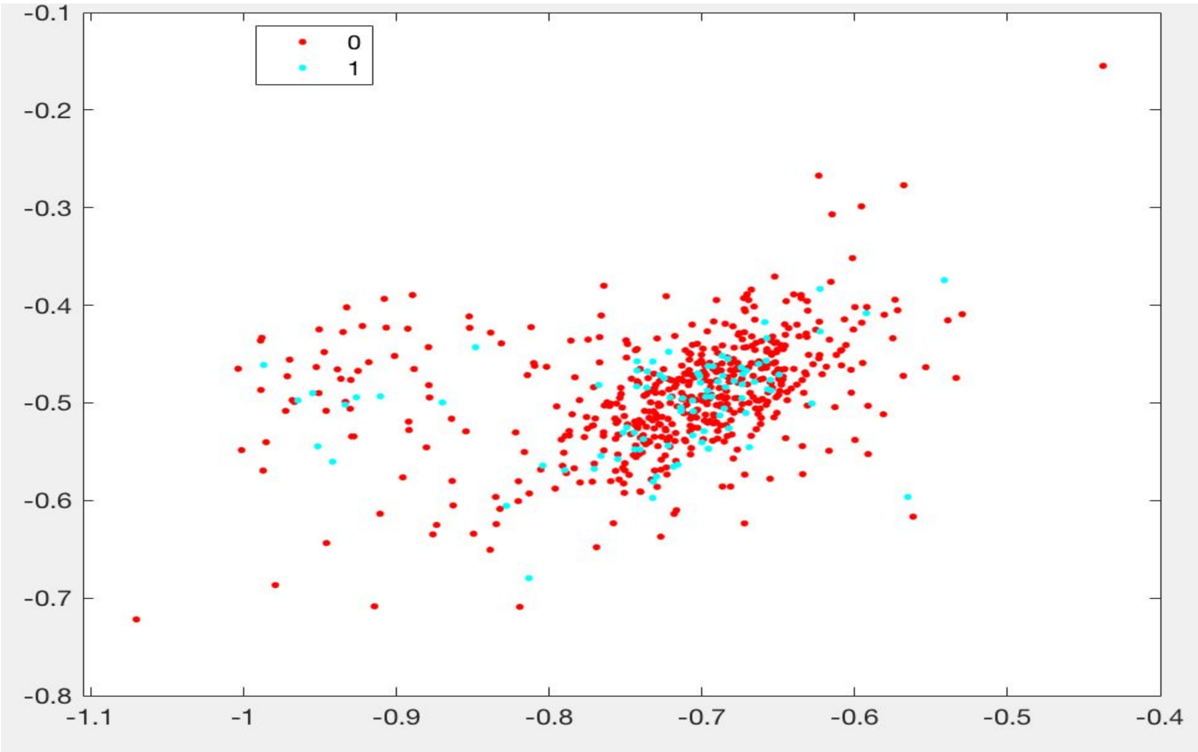
The following are six 2D scatter plots for two most important features of each emotion:
(x axis is the most important feature, y axis is the second most important feature)



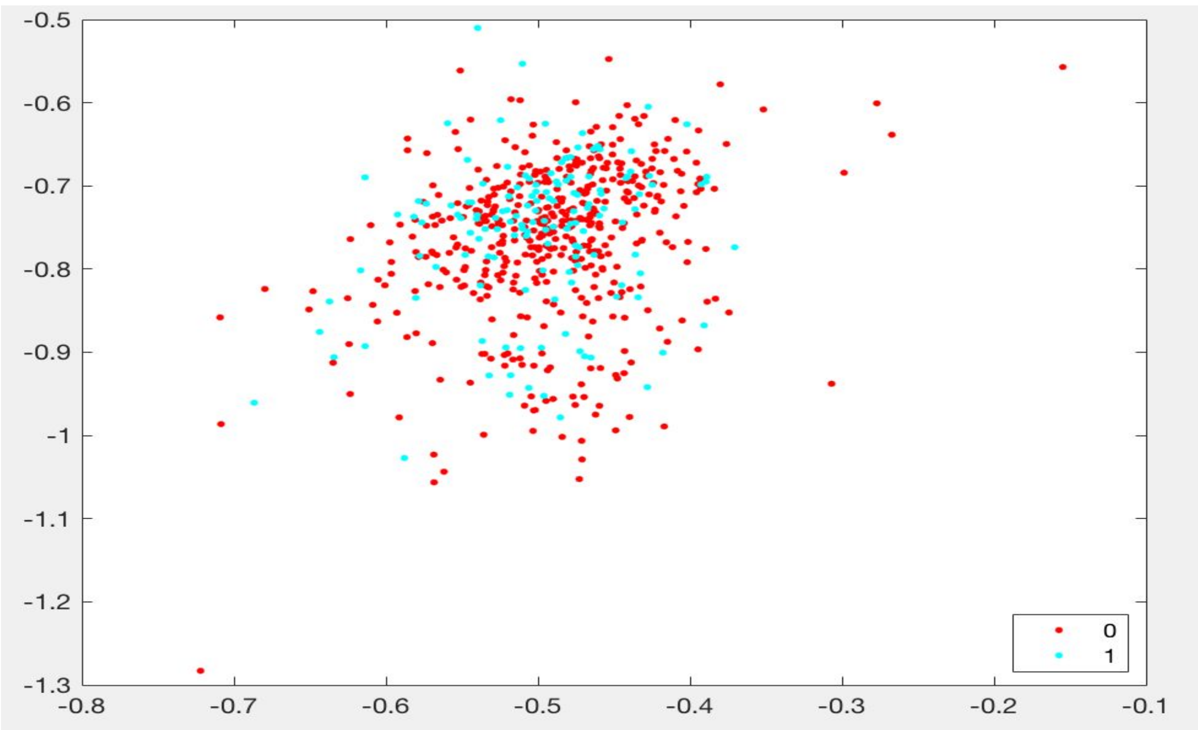
emotion anger (label 1)



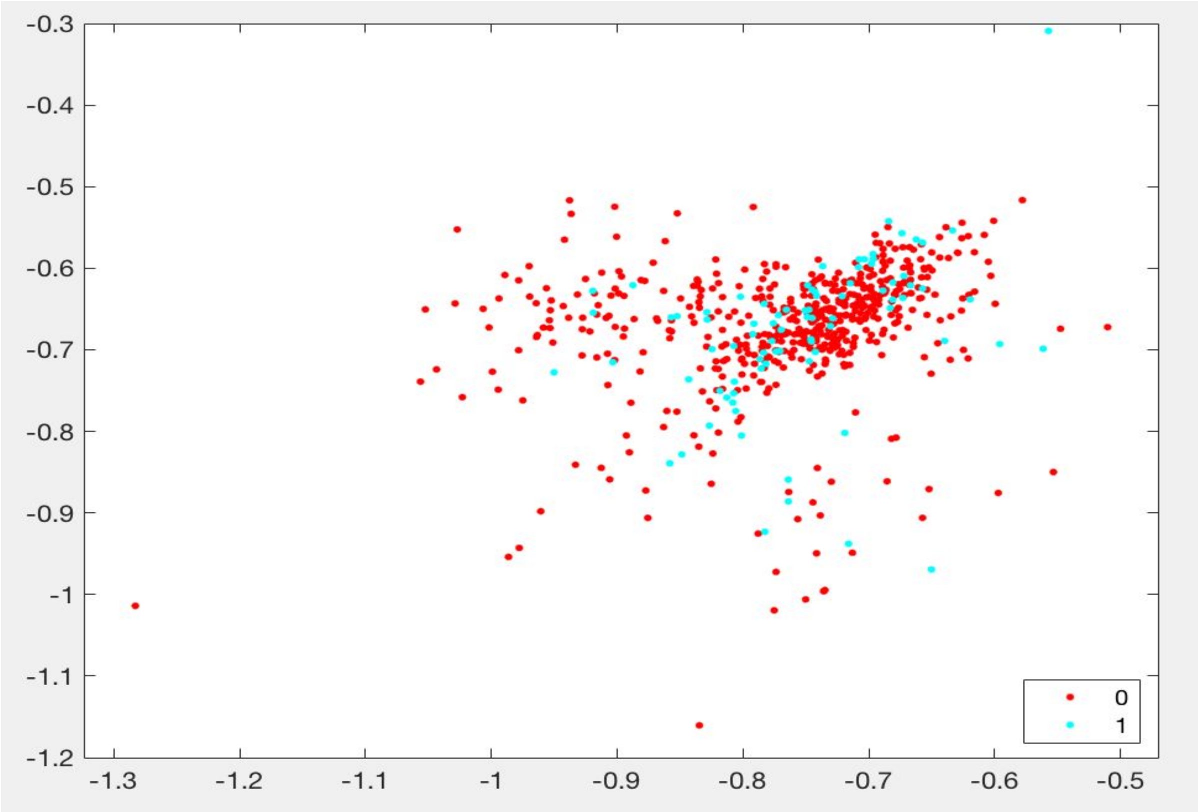
emotion disgust (label 2)



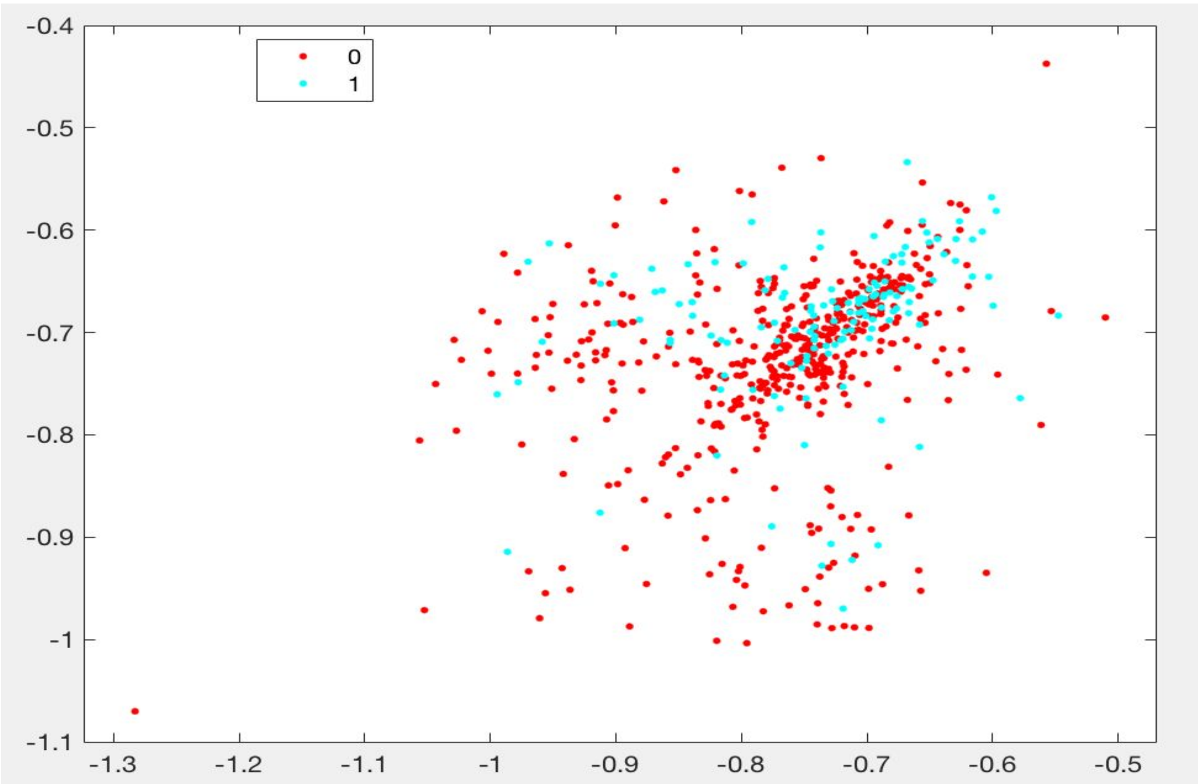
emotion fear (label 3)



emotion happiness (label 4)



emotion sadness (label 5)

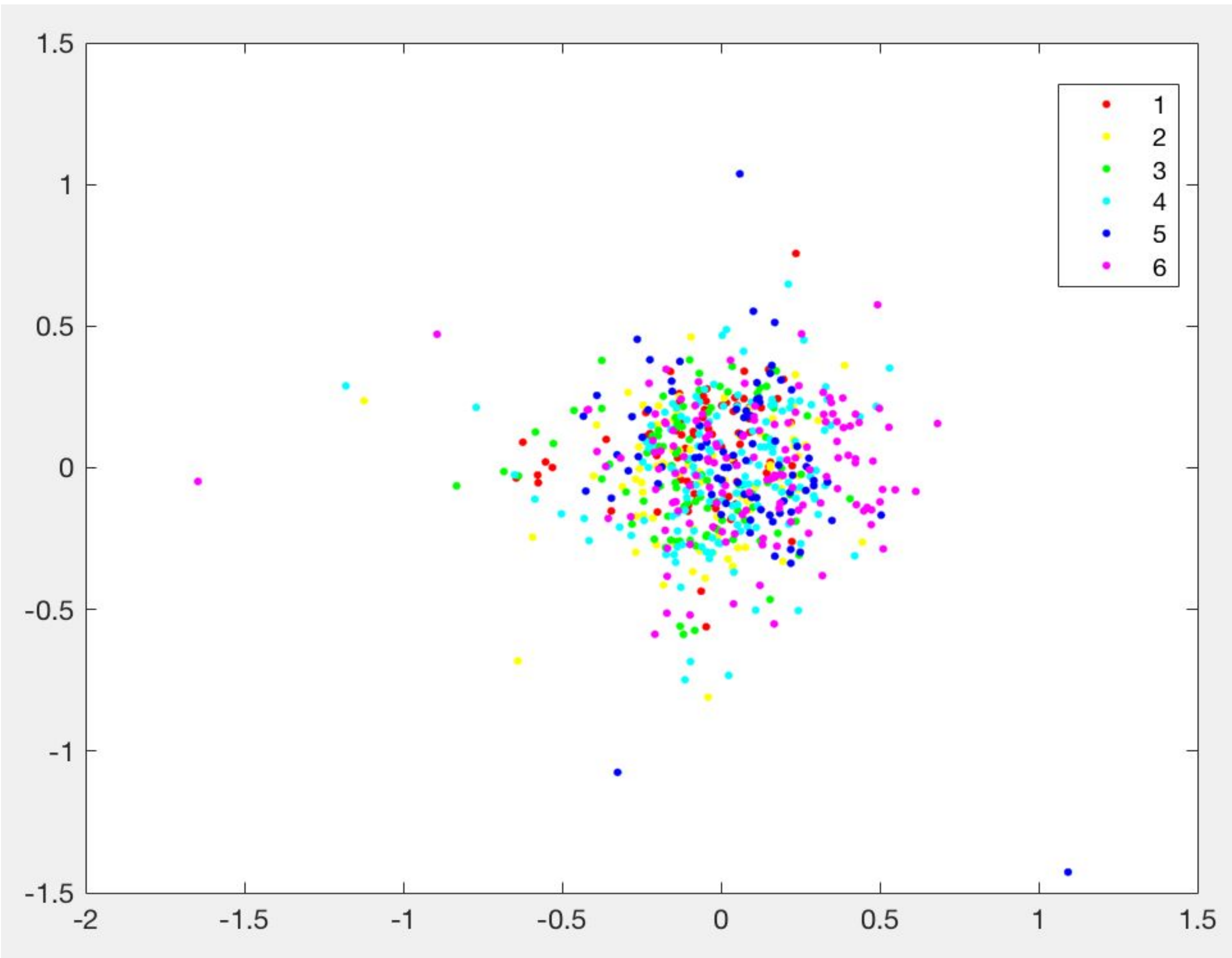


emotion surprise (label 6)

3.2 Principal Component Analysis

In matlab, Principal Component Analysis is implemented by a built-in fuction called 'pca'. To use the pca fuction, the group just wrote the script '[coeff, score, latent] = pca(x)' in the file 'cw3_pca.m'. First variable [coeff] is the principal components coefficient, the second variable [score] returns the eigenvalues which are the transformed input data, and the third variable [latent] is relative variance for each PC. As the eigenvalues are ordered in a descending order, the first two PCs are actually the first two coloums in [score]. According to the code 'cumsum(latent)./sum(latent)', the result of this code shows the cumlative sum on how much percertange of variance the number of new PCs have covered. The reason of applying this is that the eigenvalues are proportional to explain the variance. In this question, the 43th value of 'cumsum(latent)./sum(latent)' is 0.9508 which is greater than 0.9500, so this means the group should choose the first 42 PCs to cover 95% of the variance.

The following is a scatter plot of first two PCs.
(1-6 has different color and represents for 6 different emotions)
(X axis is the first PC, Y axis is the second PC)



3.3 Compare Performance

After applying 10-fold cross validation to decision tree with the data has been processed by CFS method, six trees in total are generated: one for each emotion. A test sample must be passed to all six trees before it can be classified. This can sometimes cause ambiguity: a sample may utimately have 0 or more than 1 class label which makes labeling it a hard choice. The method used here is the samples that have 0 class according to trees will be randomly assign a label among six emotions. Those have more than one class will make random choices between trees that accept them. These two improvements are implemented according to Assignment 2 feedback.

The confusion matrix and average recall, precision, F_1 measure indicates the performance of the decision tree. In the Assignment 3, it also indicates which dimensionality reduction method perform better. From the tables below, it is clear that the CFS method performs better than PCA method. In CFS confusion matrix, the sum of true positives is 287, while in the PCA confusion matrix, the sum of true positives is only 228. As for the recall, precision and F1 measure chart, nearly every value in CFS chart is greater than PCA chart. To conclude, the decision trees which applied CFS method have a better classification rate and overall a better performance.

CFS Confusion Matrix

	Output 1	Output 2	Output 3	Output 4	Output 5	Output 6
Target 1	31	16	6	9	6	6
Target 2	16	37	5	9	7	10
Target 3	13	20	31	8	6	12
Target 4	19	19	13	83	3	6
Target 5	19	15	6	9	28	7
Target 6	21	18	7	11	3	77

PCA Confusion Matrix

	Output 1	Output 2	Output 3	Output 4	Output 5	Output 6
Target 1	28	18	6	6	10	6
Target 2	17	24	12	7	10	14
Target 3	17	14	32	9	10	8
Target 4	31	23	13	60	7	9
Target 5	15	15	9	6	31	8
Target 6	32	32	5	8	7	53

CFS average recall, precision and F1 measure

	target 1	target 2	target 3	target 4	target 5	target 6
recall	41.89%	44.05%	34.44%	58.04%	33.33%	56.20%
precision	26.05%	29.60%	45.59%	64.34%	52.83%	65.25%
F1 measure	32.12%	35.41%	39.24%	61.03%	40.88%	60.39%

PCA average recall, precision and F1 measure

	Target 1	Target 2	Target 3	Target 4	Target 5	Target 6
recall	37.84%	28.57%	35.56%	41.96%	36.90%	38.69%
precision	20.00%	19.05%	41.56%	62.50%	41.33%	54.08%
F1 measure	26.17%	22.86%	38.32%	50.21%	38.99%	45.11%

3.4 Addition Questions

Question 1: The group implemeted the CFS method 6 times, each emotion was applied CFS individually and resulted into 6 different set of features each. From the formula of CFS, it is clear that the purpose of CFS is to make sure the selected amount of features are correlate strongly with the targets but correlate weakly with the already selected features. So both selected features and targets are affecting the final result in CFS method. If the group just applied the CFS once and then used it to train 6 trees at one time, the classification rate will be significantly lower than the result of applying it 6 times. In this case, the result is selected features based on the correlation with 6 targets (1-6) but not only focused on correlation with one emotion. This leads the imprecise calculation and lower the classification rate.
However, PCA is only needed to apply once. The PCA method only needs the input data (x) but not the targets (y), it transformed the input data into principal componets(PCs) and tried to find the best PCs by minimizing the reconstruction error and maximazing the variance of the projected data. Hence, no matter how many times the group applied the PCA methods, as long as the input data remains the same, the PCA result will not change.

Question 2: After implementing the PCA method, all features are converted into PC, and the theory behind it is the calculatation of the eigenvalue decomposition of input data covariance. In this question, 132 facial point features are converted to 132 PCs and these PCs actually explain the variance of the data and the new 132 PCs does not reperesent attribute number of the facial point features before applying PCA. The apporipate way to use PCA should firstly know how much variance the data is needed to be covered, then it can find the exactly amount number of PCs to satisfy the needs. When applying the matlab built-in function ‘pca’, it returns the variable [score] which are converted data and ordered by descending order of eigenvalues. It is possible to find the PC which accounts for the most variance but from my point of view, it is not the same as most informative feature. So it is impossible to infer what feature are most informative.

Question 3: As what has already been explained in the section 3.2, the group used a calculation method, ‘cumsum(latent)./sum(latent)’ in the code to analyse how latent variables affect the variance in PCA. The result of this code shows the cumlative sum on how much percertange of variance the number of new PCs have covered. The reason for applying this is that the eigenvalues are proportional to explain the variance. In this question, the 43th value of ‘cumsum(latent)./sum(latent)’ is 0.9508 which is greater than 0.9500, so this means the group should choose the first 42 PCs to cover 95% of the variance.

4. Conclusion

In the Assignment 3, the group implented CFS and PCA method to achieve dimensionality reduction. The results of confusion matrix and classification rate showed that the CFS method performs better than the PCA method for the given facial point data problem through binary decision tree.