Project Assignment 3 Report

DATA MINING

CSE 572: Spring 2018

Submitted to:

Professor Ayan Banerjee
Ira A. Fulton School of Engineering
Arizona State University

Submitted by(GROUP 23):

Abhinethri Tumkur Umesh (atumkuru@asu.edu)
Aishwarya Mohan (amohan34@asu.edu)
Jay Patel (jkpatel5@asu.edu)
Madhavi Latha Bodeddula (mbodeddu@asu.edu)
Vineesha Kasam (vkasam@asu.edu)
April 12, 2018

1. General Procedure

Based on the results of the previous task of dimensionality reduction using PCA, we have decided to pick the top four features from six, for better classification after normalising the sensor values. We only took 10 groups into account for dimensionality reduction. The new feature matrix that was obtained by multiplying the eigenvector with the old feature matrix was further used for classification. We developed a training set for each of the actions from groups of 10 users such that 60% of data per action from each team is taken to build the training set and similarly 40% of data is taken to build the test set. Testing is implemented on a user dependent basis for each of the 10 groups separately.

We have attempted to resolve the **class imbalance problem** for proper training of data. For example, for about versus non-about classification, the number of non-about instances were overpowering. Hence we decided to pick only 3 random instances from each of the non-about gestures thereby reducing the number of non-about instances for training.

2. Performance metrics

Before discussing the machine learning algorithms, it is required to know the parameters which are used to measure the performance of machine learning algorithms and how to visualize them.

2.1 Confusion Matrix

This matrix which is also called as error matrix is used to represent the values obtained on test data after predictions, from which the performance of a model can be evaluated.

Below is the sample confusion matrix:

N= Number of Observations	Predicted -> No	Predicted -> Yes	
Actual -> Yes	True Positive (TP)	False Positive (FP)	TP + FP
Actual -> No	False Negative (FN)	True Negative(TN)	FN+TP
	TP+FN	FP+TP	

True Positive (TP): It represents the total number of positive classes which have been classified accurately i.e. the class is '+ve' and it has been classified as '+ve' by the model.

True Negative (TN): It represents the total number of negative classes which have been classified accurately i.e. the class is '-ve' and it has been classified as '-ve' by the model.

False Positive (FP): The class is '-ve' but classified as '+ve' by the model.

False Negative (FN): The class is '+ve' but classified as '-ve' by the model.

2.2 Precision, Recall, F1 and Accuracy

For each action, precision, recall, F1 score and accuracy have been calculated for the interpretation of performance measure using True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN).

2.2.1 Precision

It is the ratio of correctly predicted '+ve' classes to total number of classes which are classified as '+ve' classes. If the value of this performance metric is high then it shows that the number of classes which are falsely classified as '+ve' classes is less.

Precision = TP / (TP + FP)

2.2.2 Recall

It is the ratio of correctly predicted '+ve' classes to all the actual '+ve' classes. It tells how well the model was able to classify the positive classes.

Recall = TP / (TP + FN)

2.2.3 F1 Score

It is weighted average of Precision and Recall. It is difficult to understand the concept from the definition but it is helpful when class distribution is uneven.

F1= (2 * Precision * Recall) / Precision + Recall

2.2.4 Accuracy

It is the ratio of correctly classified observations to all the observations and it is the simplest performance measure. Accuracy helps to know whether model is good as it is an intuitive parameter.

Accuracy= (TP + TN) / (TP + FP + FN + TN)

Thus, F1 score helps to evaluate the model better when FP and FN have similar cost but if the costs are different then accuracy helps in better evaluation.

3. Implementation of Machine learning algorithms

Implemented the following machine learning algorithms to classify a particular action among 10 different actions (About, And, Can, Cop, Deaf, Decide, Hearing, Father, Find and Go Out):

- Decision Tree
- Support Vector Machine
- Neural Network

3.1 Decision Tree

Decision tree is built top-down from root node to break down dataset into smaller subsets. Each internal node of decision tree represents a "test" on an attribute and each leaf node represents the classification or decision. In the assignment, we have constructed Decision Tree using matlab function **fitctree**.

Input to fitctree for classifying individual actions:

Training and Test Data:

Dimension of Training data: 480 x 4

Dimension of Test Data: 35 x 4(User specific dimension per group for each action)

Model Parameters:

Two parameters are passed: i) PruneCriterion: impurity ii) SplitCriterion: deviance

Model Details:

Number of branch nodes: 2 Number of leaf nodes: 3

Decision tree created properties for the first user: Decision cut value for first node: 2.5455e+03 Decision cut value for second node: 1.4238E+03

Few snippets from the code:

% Training the data

Total_Train_data = table(dataArray{1:end-1}, 'VariableNames', {'VarName1', 'VarName2', 'VarName3', 'VarName4', 'VarName5'});
Total_Train_data_svm = Total_Train_data{1:size(Total_Train_data,1),1:4}; % without last column
result = zeros(size(Total_Train_data,1),1);

result(1:120,1)=1;

% Call fitcsvm to train SVM model.

 $decision_model = fitctree(Total_Train_data_svm,result,'PruneCriterion','impurity','SplitCriterion','deviance');$

USER SPECIFIC TEST RESULTS FOR DECISION TREE

3.1.1 About

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.375	0.35294	68.571
2	0.8	1	0.88889	94.286

3	0.2	0.125	0.15385	68.571
4	0.5	0.125	0.2	77.143
5	0.33333	0.375	0.35294	68.571
6	0.33333	0.625	0.43478	62.857
7	0.54545	0.75	0.63158	80
8	0.66667	0.5	0.57143	82.857
9	1	0.625	0.76923	91.429
10	1	0.625	0.76923	91.429

3.1.2 And:

Group no.	Precision	Recall	f1	Accuracy
1	0.2121	0.225	0.26139	72.134
2	0.2	0.125	0.15385	68.571
3	0.5	0.125	0.2	77.143
4	0.57143	0.5	0.53333	80
5	0.5	0.125	0.2	77.143
6	0.25	0.375	0.3	60
7	0.33333	0.375	0.35294	68.571
8	0.25	0.375	0.3	60
9	0.33333	0.125	0.18182	74.286
10	0.8	1	0.88889	94.286

3.1.3 Can:

Group no.	Precision	Recall	f1	Accuracy
1	0.28571	0.25	0.26667	68.571
2	0.57143	0.5	0.53333	80
3	0.33333	0.25	0.28571	71.429

4	0.2	0.125	0.15385	68.571
5	0.4	0.25	0.30769	74.286
6	0.21429	0.375	0.27273	54.286
7	0.33333	0.625	0.43478	62.857
8	0.28571	0.25	0.26667	68.571
9	0.33333	0.125	0.18182	74.286
10	1	0.875	0.93333	97.143

3.1.4 Cop:

Group no.	Precision	Recall	f1	Accuracy
1	0.4	0.25	0.30769	74.286
2	0.55556	0.625	0.58824	80
3	0.28571	0.25	0.26667	68.571
4	0.42857	0.375	0.4	74.286
5	0.2	0.125	0.15385	68.571
6	0.55556	0.625	0.58824	80
7	1	0.875	0.93333	97.143
8	0.4	0.5	0.44444	71.429
9	0.25	0.25	0.25	65.714
10	0.33333	0.375	0.35294	68.571

3.1.5 Deaf:

Group no.	Precision	Recall	f1	Accuracy
1	0.66667	0.5	0.57143	82.857
2	0.625	0.625	0.625	82.857
3	1	0.25	0.4	82.857

4	0.25	0.125	0.16667	71.429
5	0.14286	0.125	0.13333	62.857
6	0.625	0.625	0.625	82.857
7	0.66667	0.75	0.70588	85.714
8	0.66667	0.5	0.57143	82.857
9	0.57143	0.5	0.53333	80
10	0.55556	0.625	0.58824	80

3.1.6 Decide:

Group no.	Precision	Recall	f1	Accuracy
1	0.66667	0.25	0.36364	80
2	0.66667	0.25	0.36364	80
3	0.5	0.375	0.42857	77.143
4	0.33333	0.375	0.35294	68.571
5	0.4	0.25	0.30769	74.286
6	0.4	0.25	0.30769	74.286
7	0.55556	0.625	0.58824	80
8	0.44444	0.5	0.47059	74.286
9	0.33333	0.25	0.28571	71.429
10	0.42857	0.375	0.4	74.286

3.1.7 Father:

Group no.	Precision	Recall	f1	Accuracy
1	0.375	0.375	0.375	71.429
2	0.33333	0.25	0.28571	71.429
3	0.42857	0.375	0.4	74.286

4	0.33333	0.125	0.18182	74.286
5	0.5	0.375	0.42857	77.143
6	0.25	0.375	0.3	60
7	0.4375	0.875	0.58333	71.429
8	0.25	0.375	0.3	60
9	0.375	0.375	0.375	71.429
10	0.27273	0.375	0.31579	62.857

3.1.8 Find:

Group no.	Precision	Recall	f1	Accuracy
1	0.66667	0.5	0.57143	82.857
2	0.55556	0.625	0.58824	80
3	0.375	0.375	0.375	71.429
4	0.66667	0.25	0.36364	80
5	0.66667	0.5	0.57143	82.857
6	0.5	0.5	0.5	77.143
7	0.61538	1	0.7619	85.714
8	0.4	0.5	0.44444	71.429
9	0.66667	0.5	0.57143	82.857
10	0.4	0.5	0.44444	71.429

3.1.9 GoOut:

Group no.	Precision	Recall	f1	Accuracy
1	0.63636	0.875	0.73684	85.714
2	0.5	0.375	0.42857	77.143
3	0.33333	0.125	0.18182	74.286

4	0.25	0.125	0.16667	71.429
5	0.6	0.75	0.66667	82.857
6	0.5	0.5	0.5	77.143
7	0.55556	0.625	0.58824	80
8	0.375	0.375	0.375	71.429
9	0.55556	0.625	0.58824	80
10	0.6	0.75	0.66667	82.857

3.1.10 Hearing:

Group no.	Precision	Recall	f1	Accuracy
1	0.42857	0.375	0.4	74.286
2	0.5	0.25	0.33333	77.143
3	0.5	0.375	0.42857	77.143
4	0.375	0.375	0.375	71.429
5	0.5	0.375	0.42857	77.143
6	0.55556	0.625	0.58824	80
7	0.55556	0.625	0.58824	80
8	0.3	0.375	0.33333	65.714
9	0.6	0.375	0.46154	80
10	0.66667	0.5	0.57143	82.857

3.2 Support Vector Machine

This is a supervised learning algorithm to construct the model to classify unlabeled data. In the assignment, we have constructed SVM model using matlab function **fitcsvm** to classify an action from rest of the other actions.

Input to fitcsvm for every action:

Training and Test Data:

Dimension of Training data: 480 x 4

Dimension of Test Data: 35 x 4 (User Specific - per action per group)

Model Parameters:

i) Standardize: True

It standardizes the input predictors before the training begins.

ii) 'Kernel Function': 'Gaussian'

To classify one particular class using Gaussian Kernel.

iii) BoxConstraint: 1

To specify the pair consisting of BoxConstraint and positive scalar.

iv) outlier fraction: [0,1)

To specify the expected outliers present in training data.

We have trained the model with the above model parameters and the test data has been fed into the trained model to label the unknown records. The prediction output has been classified based on the action (For example, the predicted values for all the actual 'About' classes are grouped together and remaining all other actions as another group) and this is used to calculate TP, FP, TN and FN. After calculating the above parameters, four performance metrics i.e. Precision, Recall, Fl score and Accuracy are calculated.

Snippets from the code:

% Training the data

Total_Train_data=table(dataArray{1:end-1}, 'VariableNames', {'VarName1', 'VarName2', 'VarName3', 'VarName4', 'VarName5'});
Total_Train_data_svm = Total_Train_data{1:size(Total_Train_data,1),1:4};

% without last column

result = zeros(size(Total_Train_data,1),1);

result(1:120,1)=1;

% Call fitcsvm to train SVM model.

svm_model=

 $fitcsvm(Total_Train_data_svm,result,'standardize',true,'KernelFunction','gaussian','BoxConstraint',1,'OutlierFraction',0.10);$

USER SPECIFIC TEST RESULTS FOR SVM:

3.2.1 About:

Group no.	Precision	Recall	f1	Accuracy
1	0.25	0.25	0.25	65.714
2	0.57143	0.5	0.53333	80
3	0.66667	0.25	0.36364	80

4	0.767	0.5	0.6126	77.143
5	0.856	0.625	0.2464	77.143
6	0.623	0.75	0.675	68.571
7	0.875	0.875	0.875	94.286
8	0.8	0.5	0.61538	85.714
9	1	0.75	0.85714	94.286
10	1	1	1	100

3.2.2 And:

Group no.	Precision	Recall	f1	Accuracy
1	0.71429	0.625	0.66667	77.143
2	1	0.25	0.4	82.857
3	1	0.25	0.4	82.857
4	0.436	0.25	0.567	77.143
5	0.5254	0.5	0.42857	77.143
6	0.33333	0.25	0.28571	71.429
7	0.57143	0.5	0.53333	80
8	0.5	0.375	0.42857	77.143
9	1	0.25	0.4	82.857
10	0.71429	0.625	0.66667	85.714

3.2.3 Can:

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.125	0.18182	74.286
2	0.83333	0.625	0.71429	88.571
3	1	0.25	0.4	82.857
4	0.8	0.25	0.33333	77.143

5	1	0.125	0.22222	80
6	0.33333	0.125	0.18182	74.286
7	1	0.125	0.22222	80
8	0.8	0.5	0.61538	85.714
9	1	1	1	100
10	1	0.75	0.85714	94.286

3.2.4 Cop:

Group no.	Precision	Recall	f1	Accuracy
1	0.25	0.125	0.16667	71.429
2	1	0.125	0.22222	80
3	0.33333	0.125	0.18182	68.571
4	1	0.125	0.22222	68.571
5	0.8	0.5	0.61538	65.714
6	1	0.75	0.85714	94.286
7	1	1	1	100
8	0.8	0.5	0.61538	85.714
9	0.33333	0.125	0.18182	71.429
10	0.33333	0.125	0.18182	68.571

3.2.5 Deaf:

Group no.	Precision	Recall	f1	Accuracy
1	1	1	1	100
2	1	1	1	100
3	1	0.75	0.85714	94.286
4	1	0.125	0.22222	80

5	0.5	0.375	0.42857	77.143
6	1	0.875	0.93333	97.143
7	1	0.875	0.93333	97.143
8	1	0.5	0.66667	88.571
9	0.5	0.375	0.42857	77.143
10	0.5	0.375	0.42857	77.143

3.2.6 Decide:

Group no.	Precision	Recall	f1	Accuracy
1	0.28571	0.25	0.26667	68.571
2	0.25	0.125	0.16667	71.429
3	0.5	0.125	0.33333	68.571
4	0.28571	0.25	0.26667	68.571
5	1	0.5	0.66667	88.571
6	1	0.125	0.22222	80
7	0.5	0.375	0.42857	77.143
8	0.28571	0.25	0.26667	68.571
9	0.5	0.375	0.42857	77.143
10	0.5	0.25	0.35678	71.429

3.2.7 Hearing:

Group no.	Precision	Recall	f1	Accuracy
1	1	1	1	100
2	1	0.75	0.85714	94.286
3	1	0.375	0.54545	85.714
4	0.5	0.375	0.42857	77.143

5	0.5	0.375	0.42857	77.143
6	0.72727	1	0.84211	91.429
7	0.8	1	0.88889	94.286
8	0.66667	0.75	0.70588	85.714
9	1	0.5	0.66667	88.571
10	1	0.625	0.76923	91.429

3.2.8 Father:

Group no.	Precision	Recall	f1	Accuracy
1	1	0.125	0.22222	80
2	0.5	0.25	0.35678	71.429
3	0.5	0.125	0.2	77.143
4	0.5	0.375	0.42857	77.143
5	1	0.5	0.66667	88.571
6	0.625	0.625	0.625	82.857
7	0.71429	0.625	0.66667	85.714
8	0.66667	0.5	0.57143	82.857
9	0.5	0.375	0.42857	77.143
10	0.5	0.25	0.35678	71.429

3.2.9 Find:

Group no.	Precision	Recall	f1	Accuracy
1	0.66667	0.5	0.57143	82.857
2	0.75	0.75	0.75	88.571
3	1	0.25	0.4	82.857
4	1	0.25	0.4	82.857

5	0.5	0.375	0.42857	77.143
6	0.83333	0.625	0.71429	88.571
7	0.66667	0.5	0.57143	82.857
8	1	0.75	0.85714	94.286
9	0.7	0.875	0.77778	88.571
10	0.8	1	0.88889	94.286

3.2.10 Go Out:

Group no.	Precision	Recall	f1	Accuracy
1	0.88889	1	0.94118	97.143
2	1	1	1	100
3	1	0.625	0.76923	91.429
4	0.5	0.375	0.42857	77.143
5	1	0.125	0.22222	80
6	1	0.75	0.85714	94.286
7	1	1	1	100
8	0.88889	1	0.94118	97.143
9	1	1	1	100
10	1	1	1	100

3.3 Neural Network

A Neural Network contains an input layer, a set of hidden layers and an output layer with varying number of neurons in each layer (input, output, hidden) depending on the application. In the assignment, we use the **nntool** of MATLAB and the output layer consists of only one neuron, which signifies the classification of each action (Eg: About or Not About).

Input to patternnet:

Patternet (Pattern recognition networks) classify the inputs given to them according to the target classes specified, which are the feed-forward networks. The input to the patternnet is the number of neurons in the hidden layer.

Training and Test Data:

Dimension of Training data: 480 x 4

Dimension of Test Data: 35 x 4 (User Specific - per action per group)

Model Parameters:

Two parameters are passed:
i) TrainingFunction: 'traingda'
ii) Performance Function: 'mse'

Model Details:

Number of neurons in input layer - 4 Number of neurons in hidden layer - 15 Number of neurons in the output layer - 1 Number of layers (input, hidden, output) - 1

Below is the snippet from the code:

% Training the data

net = patternnet(hiddenLayerSize); % pattern recognition network net.trainFcn = 'traingda'; net.performFcn = 'mse';

 $neural_model = train(net, transpose(Total_Train_data_neural), transpose(result));$

USER SPECIFIC TEST RESULTS FOR NEURAL NETWORK:

3.3.1 About:

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.25	0.28571	71.429
2	0.44444	0.5	0.47059	74.286
3	0.28571	0.25	0.26667	68.571
4	0.14286	0.125	0.13333	62.857

5	0.44444	0.5	0.47059	74.286
6	0.14286	0.125	0.13333	62.857
7	0.66667	0.75	0.70588	85.714
8	0.66667	0.5	0.57143	82.857
9	0.625	0.625	0.625	82.857
10	0.8	1	0.88889	94.286

3.3.2 And :

Group no.	Precision	Recall	f1	Accuracy
1	0.5	0.25	0.33333	77.143
2	0.6	0.375	0.46154	80
3	1	0.25	0.4	82.857
4	0.8	1	0.88889	94.286
5	0.66667	0.75	0.70588	85.714
6	0.4	0.75	0.52174	68.571
7	0.5	0.875	0.63636	77.143
8	0.5	0.25	0.33333	77.143
9	0.44444	0.5	0.47059	74.286
10	0.5	0.25	0.33333	77.143

3.3.3 Can:

Group no.	Precision	Recall	f1	Accuracy
1	0.5	0.125	0.2	77.143
2	0.28571	0.25	0.26667	68.571
3	0.5	0.25	0.33333	77.143
4	0.28571	0.25	0.26667	68.571
5	0.5	0.25	0.33333	77.143

6	0.33333	0.25	0.28571	71.429
7	0.33333	0.25	0.28571	71.429
8	0.14286	0.125	0.13333	62.857
9	0.14286	0.125	0.13333	62.857
10	0.25	0.25	0.25	65.714

3.3.4 Cop:

Group no.	Precision	Recall	f1	Accuracy
1	0.5	0.25	0.33333	77.143
2	0.28571	0.25	0.26667	68.571
3	0.25	0.125	0.16667	71.429
4	0.5	0.25	0.33333	77.143
5	0.5	0.25	0.33333	77.143
6	0.5	0.25	0.33333	77.143
7	0.66667	0.25	0.36364	80
8	0.33333	0.25	0.28571	71.429
9	0.66667	0.25	0.36364	80
10	0.66667	0.25	0.36364	80

3.3.5 Deaf:

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.25	0.28571	71.429
2	0.4	0.25	0.30769	74.286
3	1	0.125	0.22222	80
4	1	0.125	0.22222	80
5	0.5	0.125	0.2	77.143

6	0.5	0.5	0.5	77.143
7	0.55556	0.625	0.58824	80
8	0.42857	0.375	0.4	74.286
9	0.5	0.375	0.42857	77.143
10	0.33333	0.25	0.28571	71.429

3.3.6 Decide:

Group no.	Precision	Recall	f1	Accuracy
1	0.4	0.25	0.30769	74.286
2	0.66667	0.25	0.36364	80
3	0.28571	0.25	0.26667	68.571
4	0.42857	0.375	0.4	74.286
5	0.5	0.375	0.42857	77.143
6	0.5	0.375	0.42857	77.143
7	0.33333	0.25	0.28571	71.429
8	0.5	0.375	0.42857	77.143
9	0.4	0.25	0.30769	74.286
10	0.66667	0.25	0.36364	80

3.3.7 Father:

Group no.	Precision	Recall	f1	Accuracy
1	0.625	0.625	0.625	82.857
2	0.5	0.5	0.5	77.143
3	0.5	0.375	0.42857	77.143
4	0.30769	0.5	0.38095	62.857
5	0.5	0.375	0.42857	77.143

6	0.66667	0.75	0.70588	85.714
7	0.5	0.5	0.5	77.143
8	0.5	0.375	0.42857	77.143
9	0.5	0.25	0.33333	77.143
10	0.66667	0.25	0.36364	80

3.3.8 Find:

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.25	0.28571	71.429
2	0.33333	0.25	0.28571	71.429
3	0.25	0.25	0.25	65.714
4	0.30769	0.5	0.38095	62.857
5	0.18182	0.25	0.21053	57.143
6	0.5	0.5	0.5	77.143
7	0.625	0.625	0.625	82.857
8	0.5	0.375	0.42857	77.143
9	0.66667	0.25	0.36364	80
10	0.5	0.25	0.33333	77.143

3.3.9 GoOut:

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.25	0.28571	71.429
2	0.42857	0.375	0.4	74.286
3	0.4	0.25	0.30769	74.286
4	0.66667	0.25	0.36364	80
5	0.75	0.375	0.5	82.857

6	0.5	0.5	0.5	77.143
7	0.4	0.25	0.30769	74.286
8	0.5	0.375	0.42857	77.143
9	0.22222	0.25	0.23529	62.857
10	0.33333	0.25	0.28571	71.429

3.3.10 Hearing:

Group no.	Precision	Recall	f1	Accuracy
1	0.33333	0.25	0.28571	71.429
2	0.42857	0.375	0.4	74.286
3	0.5	0.25	0.33333	77.143
4	0.66667	0.25	0.36364	80
5	0.5	0.125	0.2	77.143
6	0.42857	0.375	0.4	74.286
7	0.66667	0.5	0.57143	82.857
8	0.33333	0.25	0.28571	71.429
9	0.5	0.25	0.33333	77.143
10	0.66667	0.25	0.36364	80

PERFORMANCE EVALUATION:

ABOUT: SVM performs better with an average accuracy about 80% compared to neural network and decision tree.

AND: Neural Network performs well with an average accuracy about 80% in comparison to SVM and decision tree.

CAN: SVM outperforms with an average accuracy about 85% compared to neural network and decision tree.

COP: A good performance is shown by Neural Network with an average accuracy of 75% compared to SVM and decision tree.

DEAF: A good accuracy SVM performs better with an average accuracy of 85% compared to neural network and decision tree.

DECIDE: Decision tree performs better with an average accuracy about 75% compared to SVM and neural network.

FATHER: SVM outperforms neural networks and decision tree with an average accuracy of 80%.

FIND: SVM shows good performance with an average accuracy about 85% compared to neural network and decision tree.

GO OUT: SVM outperforms with an average accuracy about 90% compared to neural network and decision tree.

HEARING: SVM shows good performance with an average accuracy about 85% compared to neural network and decision tree.