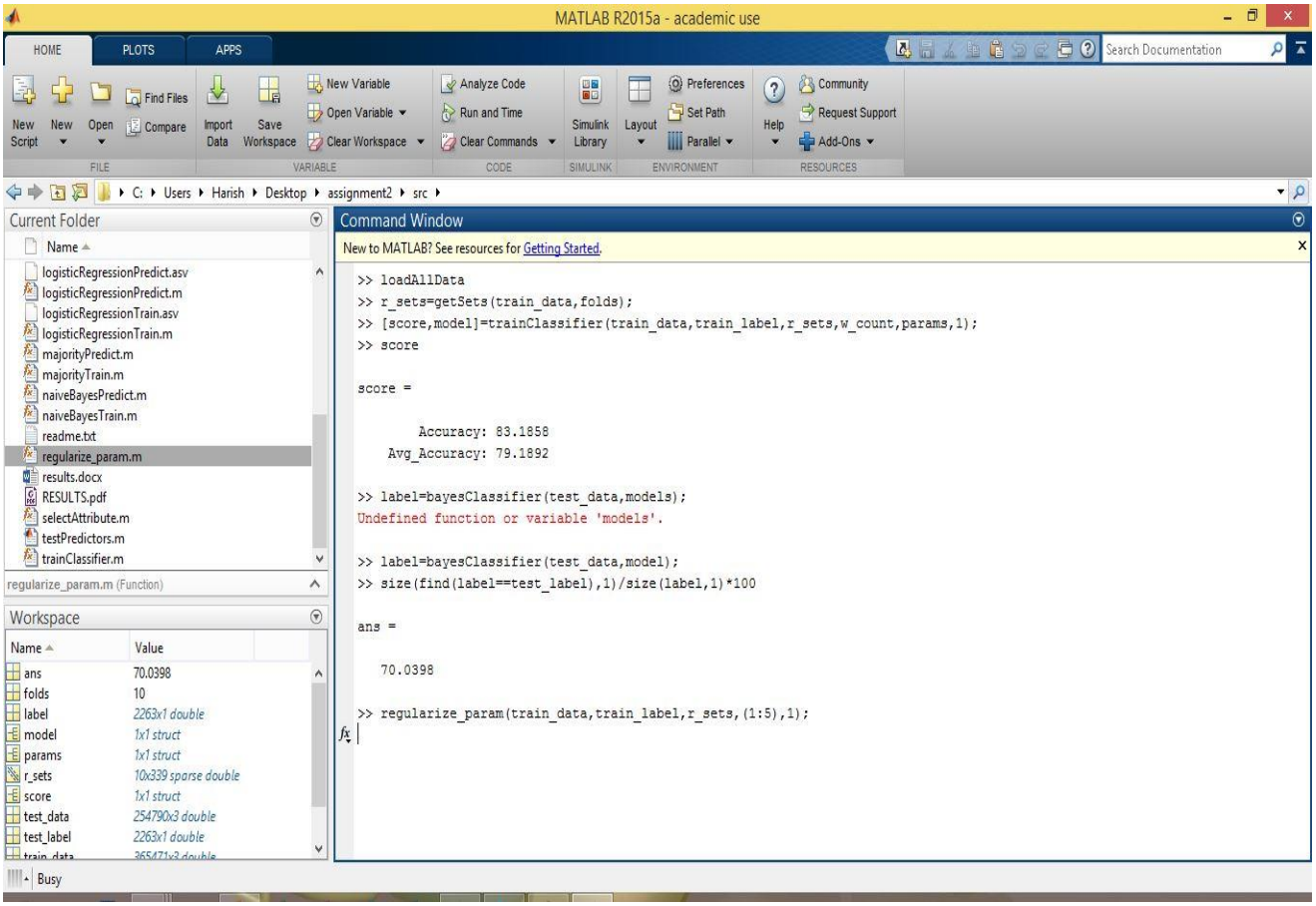
Machine Learning – ITCS 6156

Assignment – II

Results:

1. Naive Bayes Classification:



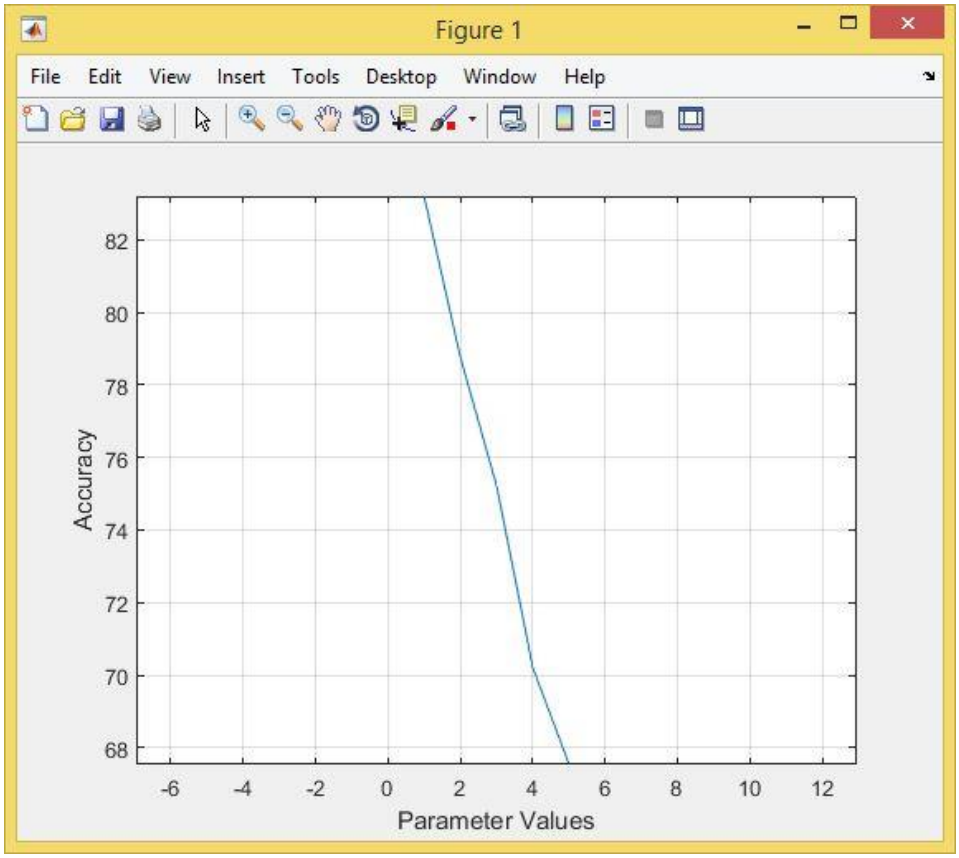


Figure 1
File Edit View Insert Tools Desktop Window Help
Accuracy
82
80
78
76
74
72
70
68
-6 -4 -2 0 2 4 6 8 10 12
Parameter Values

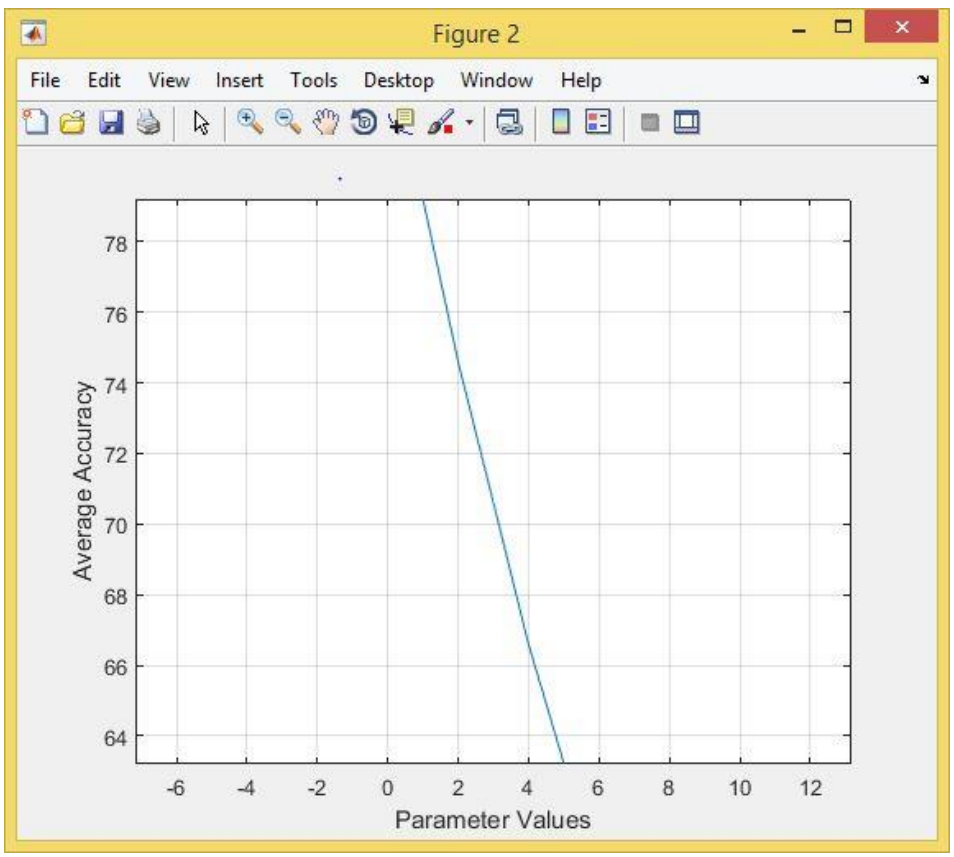
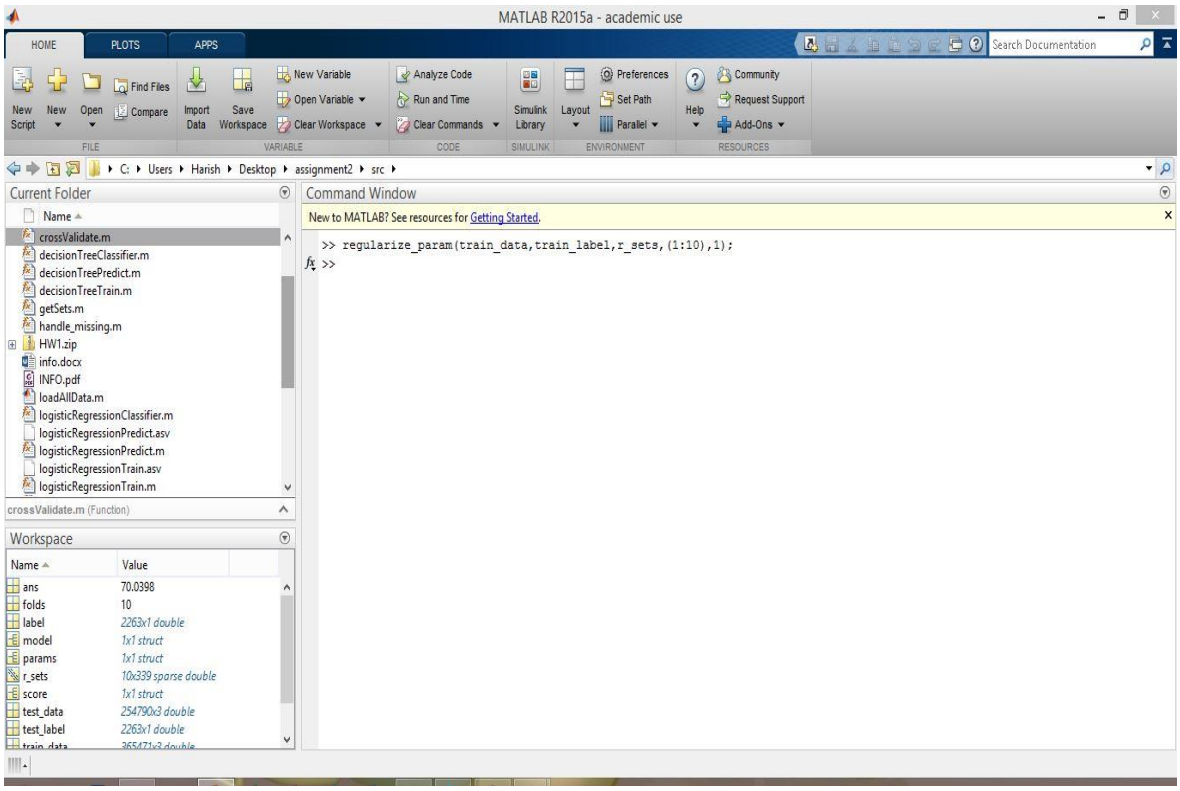


Figure 2
File Edit View Insert Tools Desktop Window Help
Average Accuracy
78
76
74
72
70
68
66
64
-6 -4 -2 0 2 4 6 8 10 12
Parameter Values


MATLAB R2015a - academic use
Current Folder
Command Window
New to MATLAB? See resources for Getting Started.
>> regularize_param(train_data,train_label,r_sets,(1:10),1);
>>
crossValidate.m
decisionTreeClassifier.m
decisionTreePredict.m
decisionTreeTrain.m
getSets.m
handle_missing.m
HW1.zip
info.docx
INFO.pdf
loadAllData.m
logisticRegressionClassifier.m
logisticRegressionPredict.asv
logisticRegressionPredict.m
logisticRegressionTrain.asv
logisticRegressionTrain.m
Workspace
ans 70.0398
folds 10
label 2263x1 double
model 1x1 struct
params 1x1 struct
r_sets 10x339 sparse double
score 1x1 struct
test_data 254790x3 double
test_label 2263x1 double

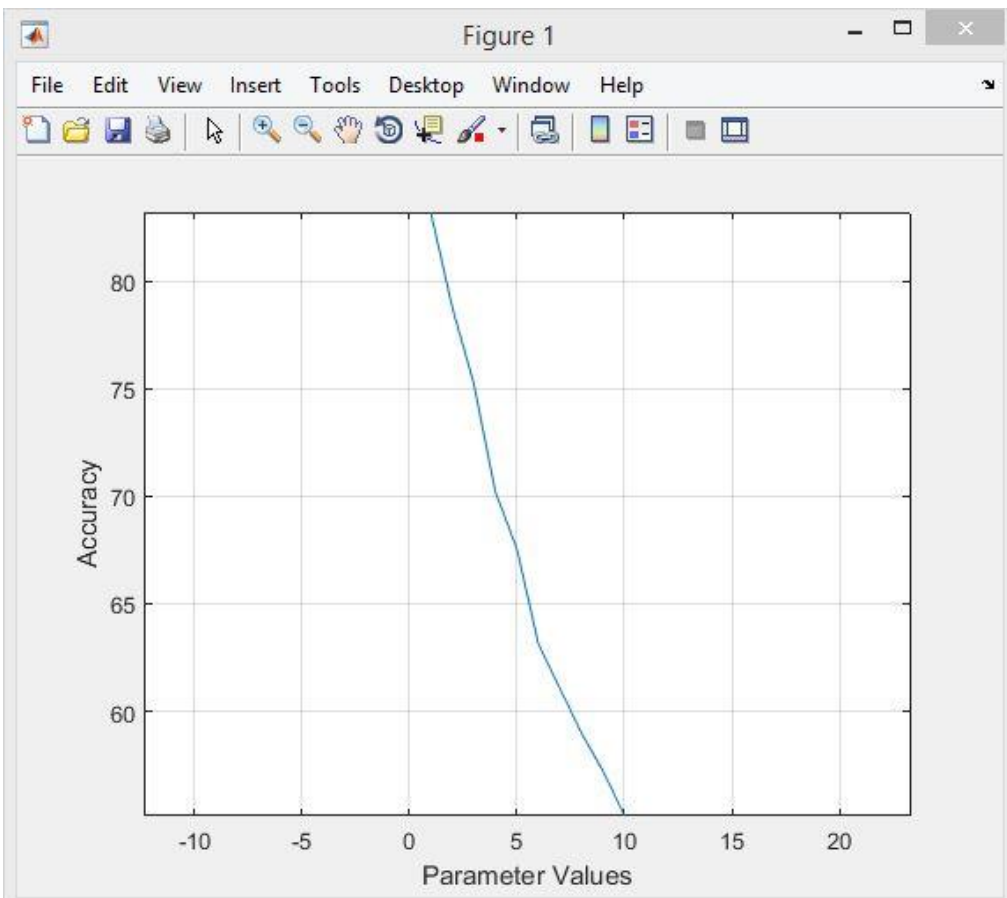
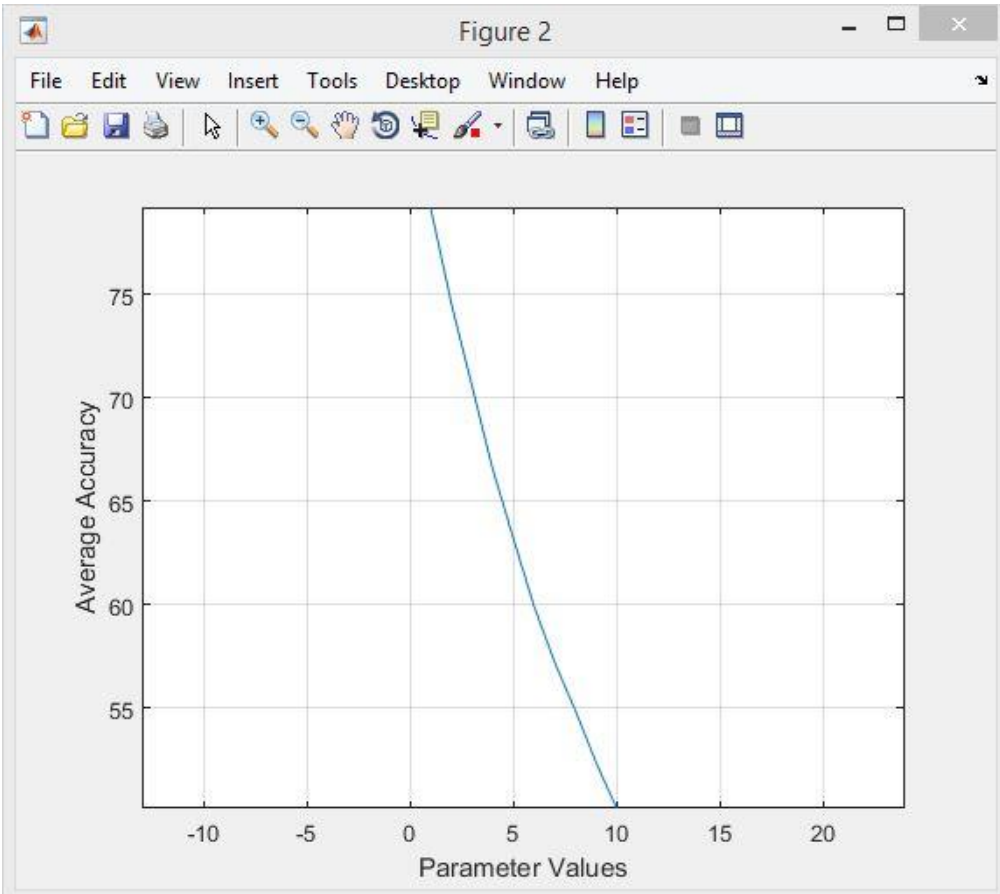


Figure 1
File Edit View Insert Tools Desktop Window Help
Accuracy
80
75
70
65
60
-10
-5
0
5
10
15
20
Parameter Values



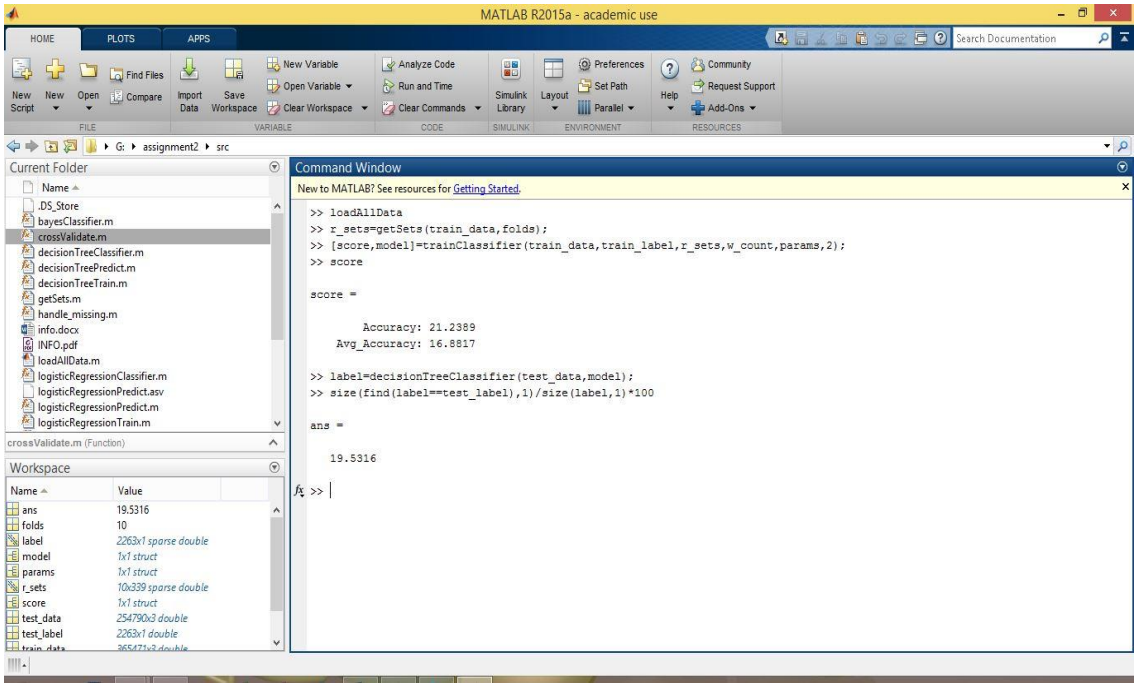
My observation for this classification was as follows: when alpha is 1 it becomes the maximum of the classifier output than the remaining values of it {0,2, 3,…}. I performed the classification on entire data set rather than considering the set consisting of only some attributes because the accuracy decreases to around 40% on training set and 37% on testing data. The above shown figures are taken when alpha=1.

2. **Decision Tree Classification:**

My observation for this classification was as follows: when depth increases the accuracy of the model also keeps increasing. Due to the performance issues I presents some raw results here after iterating the crossValidate.m for one iteration:

Depth	Accuracy (training)
10	17.2172%
20	22.4189%

The above table is formulated when training the classifier on the set with the selected word_id's. The following figure gives the result for crossValidating on the set with selective attributes.



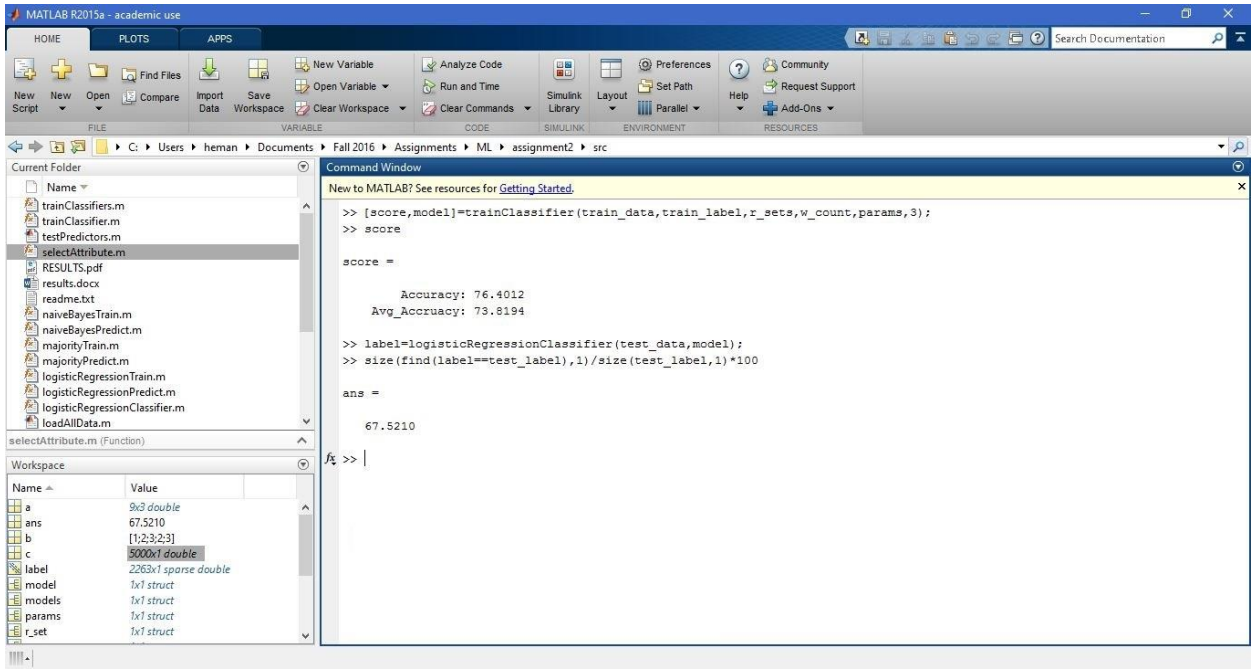
The efficiency of this classifier can be increased by choosing number of words as more than 5000 or to cross validate the classifier on entire data set, and also by handling missing values in each words in each document. However, to handle the missing values the classifier performance is too low therefore, I tried to implement it on the entire dataset to see its efficiency. Because of time issue I am unable to show you the result.

3. Logistic Regression Classification:

My observation for this classification was as follows: when running the classifier using the selected word_ids, it gives the following results:

Alpha = 10	
Iteration Limit	**Accuracy (training)**
50	41.2979%
100	41.8879%
500	42.1829%

In the above table I choose alpha as 10 such that it's the maximum alpha that suits to correctly converge the theta values i.e. when we keep >10 the algorithm jumps the limit of converging and if <10 more number of iterations are required to converge. As we can see the maximum accuracy with this approach may be limited to 50%. Therefore, I implemented the classifier on the entire dataset. The following figure is collected by classifying on the entire data set and for this context the alpha is tuned to 0.01:


MATLAB R2015a - academic use
HOME
PLOTS
APPS
Search Documentation
New Script
New
Open
Find Files
Compare
Import Data
Save Workspace
New Variable
Open Variable
Clear Workspace
Analyze Code
Run and Time
Clear Commands
Simulink Library
Layout
Preferences
Set Path
Parallel
Help
Community
Request Support
Add-Ons
FILE
VARIABLE
CODE
SIMULINK
ENVIRONMENT
RESOURCES
C: ▸ Users ▸ heman ▸ Documents ▸ Fall 2016 ▸ Assignments ▸ ML ▸ assignment2 ▸ src
Current Folder
Name
trainClassifiers.m
trainClassifier.m
testPredictors.m
selectAttribute.m
RESULTS.pdf
results.docx
readme.txt
naiveBayesTrain.m
naiveBayesPredict.m
majorityTrain.m
majorityPredict.m
logisticRegressionTrain.m
logisticRegressionPredict.m
logisticRegressionClassifier.m
loadAllData.m
selectAttribute.m (Function)
Workspace
Name
Value
a
9x3 double
ans
67.5210
b
[1;2;3;2;3]
c
5000x1 double
label
2263x1 sparse double
model
1x1 struct
models
1x1 struct
params
1x1 struct
r_set
1x1 struct
Command Window
New to MATLAB? See resources for Getting Started.
>> [score,model]=trainClassifier(train_data,train_label,r_sets,w_count,params,3);
>> score
score =
Accuracy: 76.4012
Avg_Accruacy: 73.8194
>> label=logisticRegressionClassifier(test_data,model);
>> size(find(label==test_label),1)/size(test_label,1)*100
ans =
67.5210
>>