Pattern Recognition
Department of Electrical and Information Engineering
University of Cassino and Southern Latium, Second Semester 2018
Homework Assignment 3
Assigned 30 May 2018; due 11:59 PM, 10 June 2018

In preparing my solutions, I did not look at any old homework's, copy anybody's answers or let them copy mine.

NAME: FAKRUL ISLAM TUSHAR. Signature:

Problem 3.1 [60%]: Consider the dataset contained in the file hw3data.csv available on the course website. It contains 8,000 samples coming from a two {class problem, each made of 10 numerical features and a binary label (±1). Consider the following tasks:

(a) Split the data into a training set (40%), validation set (40%) and test set (20%).
(b) Use the training set for building a v-SVM and the validation set for choosing the optimal value (i.e. the one maximizing the AUC) of v
(c) Evaluate the AUC on the test set.

Execute the steps above with 4 different SVM kernels (linear, polynomial of degree 2, RBF, sigmoid). For each test, consider at least 5 different splits and evaluate average AUC and standard deviation. Discuss the results obtained.

Answer: Support vector machines (SVM) is a supervised learning method. SVM is effective in high dimensional spaces and use support vectors for decision function which is memory efficient.

To perform the tasks asked in the problem 2.1 first need to prepare the dateset.

Dataset:

Provided hwk3data.csv data set was used in this experiment. The dataset contains 8,000 samples coming from a two-class problem, each made of 10 numerical features and a binary label (±1). So, first task was converting the level -1 to label 0. Pseudo code of making the lebel -1 shown below.

General pseudocode of making the label (-1) to 0

```
I.    Import necessary libraries
II.   Load the dataset
III.  Preprocessing dataset
IV.       Getting the label -1 to 0
V.        if the label is -1
VI.             Set the label=0
VII.      else
VIII.           Set the label=1
```

Now we need to divide the dataset into 5 subsets for the cross validation of the AUC with different hyperparameters. Dataset has 8000 samples where each class has 4000 samples. So it was an balances dataset. To divide this dataset into 5 equal subsets, each subset will contain 1600 samples of 800 from each class. Steps given bellow was been followed for making the subsets.

1. ScikitLearn function train_test_split was used for data split.
2. First data is divided into 5 equal subsets. Each contains 1600 samples.
3. Thus, subset of samples x_set1, x_set2, x_set3, x_set4, x_set5 achieved with subset of labels y_set1, y_set2, y_set3, y_set4, y_set5
4. Afterward combining this subset, we achieved five different of dataset with different combination shown in Tabel.1.

Tabel.1: Datasets

Dataset	Train	Test
Dataset-1	**Sample**: x_set2, x_set3, x_set4, x_set5 **Labels**: y_set2, y_set3, y_set4, y_set5	**Sample**: x_set1 **Labels**: y_set1
Dataset-2	**Sample**:x_set1, x_set3, x_set4, x_set5 **Labels**: y_set1, y_set3, y_set4, y_set5	**Sample**: x_set2 **Labels**: y_set2
Dataset-3	**Sample**: x_set1, x_set2, x_set4, x_set5 **Labels:** y_set1, y_set2, y_set4, y_set5	**Sample**: x_set3 **Labels**: y_set3
Dataset-4	**Sample**:x_set1, x_set2, x_set3, x_set5 **Labels:** y_set1, y_set2, y_set3, y_set5	**Sample**: x_set4 **Labels**: y_set4
Dataset-5	**Sample**: x_set1, x_set2, x_set3, x_set4 **Labels:** y_set1, y_set2, y_set3, y_set4	**Sample**: x_set5 **Labels**: y_set5

5. Then each Dataset was split into 40% training and 40% validation.

Hyperparameters is important in case of using SVM as they usually have huge impact on the results. Here we tried different value of nu and different kernels. Nu is to tune the tradeoff between overfitting and generalization. nu is upper bounded by the fraction of outliers and lower bounded by the fraction of SVs and, equals both in the limit.

Taking small values of nu leads to a large c, meaning that mis-classification have a large impact upon the objective. In other words, fitting the data is more important than heading for easier solutions. For large nus, the quadratic regularization is more important and label deviations wrt ground-truth data are more likely to occur, hence easier solutions with a usually better generalization ability are preferred. Experimental outcomes shown below.

❖ Case: Linear Kernel

Different values of nu were used with different datasets that we have shown in the earlier dataset section. Each of the dataset is divided into 40% training and validation set. 40% training data was being used to train the NuSVM and 40% validation was being used to finding the AUC. The results shown in Tabel.2, Tabel.3, Tabel.4, Tabel.5and Tabel.6.

Tabel.2: Linear Kernel using Training: DateSet-1

nu	0.005	0.01	0.03	0.05	0.1	0.3	0.5	0.6	0.8	0.9
AUC	0.54	0.63	0.76	0.79	0.70	0.80	0.78	0.81	0.77	0.76

Tabel.3: Linear Kernel using Training: DateSet-2

nu	0.005	0.01	0.03	0.05	0.1	0.3	0.5	0.6	0.8	0.9
AUC	0.78	0.66	0.61	0.21	0.50	0.31	0.79	0.82	0.77	0.76

Tabel.4: Linear Kernel using Training: DateSet-3

nu	0.005	0.01	0.03	0.05	0.1	0.3	0.5	0.6	0.8	0.9
AUC	0.67	0.75	0.30	0.64	0.72	0.79	80	0.81	0.77	0.76

Tabel.5: Linear Kernel using Training: DateSet-4

nu	0.005	0.01	0.03	0.05	0.1	0.3	0.5	0.6	0.8	0.9
AUC					0.79	0.80	0.80	0.81	0.77	0.76

Tabel.6: Linear Kernel using Training: DateSet-5

nu	0.005	0.01	0.03	0.05	0.1	0.3	0.5	0.6	0.8	0.9
AUC					0.72	0.77	0.78	0.80	0.76	0.76

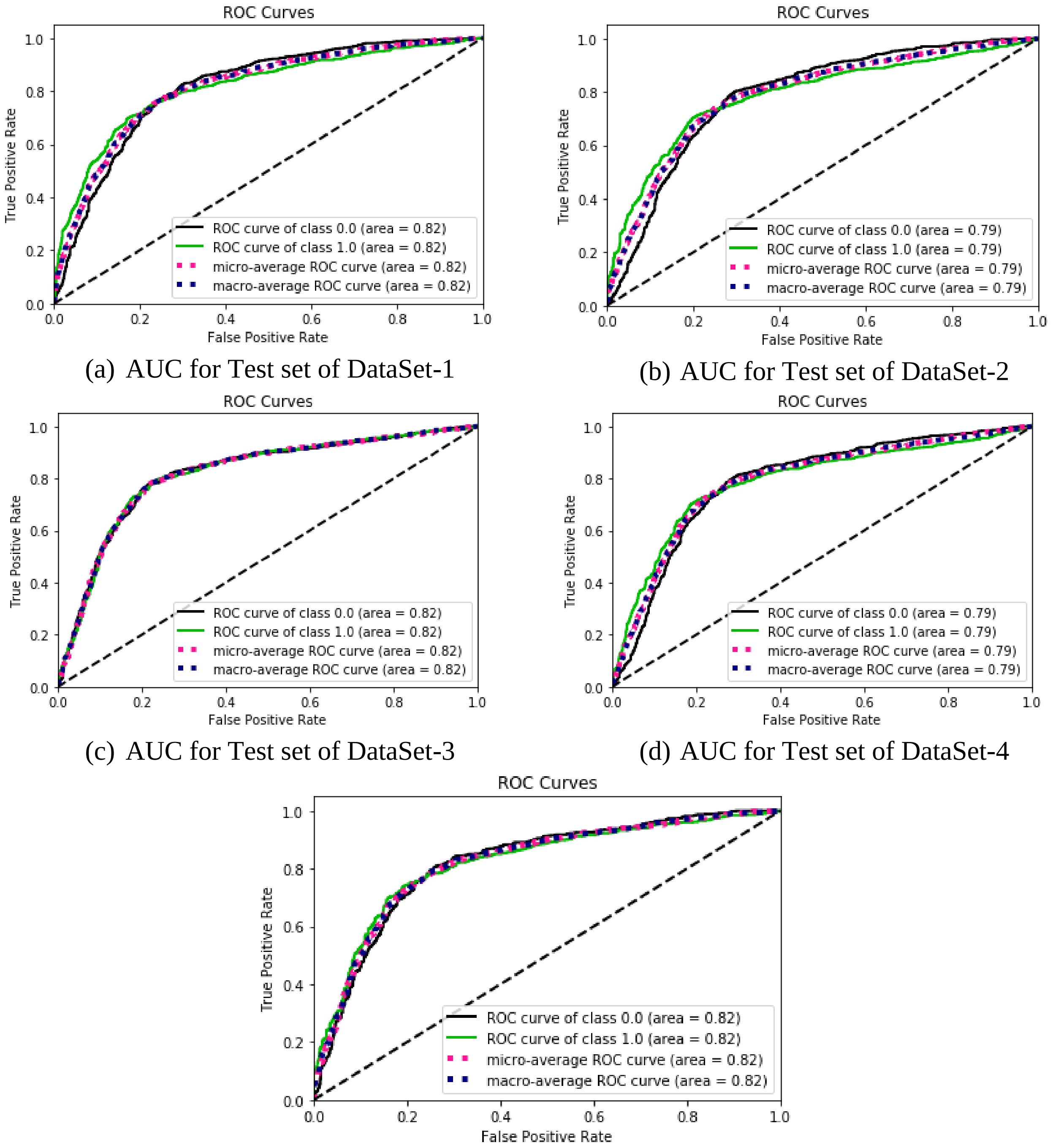
From Tabel.2, Tabel.3, Tabel.4, Tabel.5and Tabel.6 we can be seen that in the case of linear kernel the maximum AUC for validation sets is achieved when the nu is 0.6 in all the Datasets. So, nu=0.6 is been chosen as optimal value for the linear kernel.

Now using the linear kernel and nu=0.6 the NuSVM is been trained on DataSet-1, DataSet-2, DataSet-3, DataSet-4, and DataSet-5 and AUC was AUC was achieved using Test sets of the Datasets.

Tabel.7: Linear Kernel using Testing

Test No	1	2	3	4	5
AUC	0.82	0.79	0.82	0.79	0.82
Average AUC	**0.808**		**Standard Deviation**	**0.0063**	

Table. 7 shown the AUC on the test sets. The average AUC is **0.808** and the standard deviation is **0.0063**. Fig.1 shown the AUC of the test sets.



(a) AUC for Test set of DataSet-1

(b) AUC for Test set of DataSet-2

(c) AUC for Test set of DataSet-3

(d) AUC for Test set of DataSet-4

(e) AUC for Test set of DataSet-5

Fig.1: AUC for the test sets.

❖ Case: RBF Kernel

Different values of nu were used with different datasets that we have shown in the earlier dataset section. Each of the dataset is divided into 40% training and validation set. 40% training data was being used to train the NuSVM and 40% validation was being used to finding the AUC. The results shown in Tabel.8, Tabel.9, Tabel.10, Tabel.11and Tabel.12.

Tabel.8: Linear Kernel using Training: DateSet-1

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.19	0.81	0.84	0.84	0.80	0.77	0.76

Tabel.9: Linear Kernel using Training: DateSet-2

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.75	0.83	0.86	0.85	0.80	0.77	0.76

Tabel.10: Linear Kernel using Training: DateSet-3

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.71	0.79	0.84	0.84	0.80	0.77	0.76

Tabel.11: Linear Kernel using Training: DateSet-4

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.50	0.84	0.84	0.83	0.79	0.77	0.76

Tabel.12: Linear Kernel using Training: DateSet-5

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.75	0.80	0.85	0.84	0.79	0.77	0.76

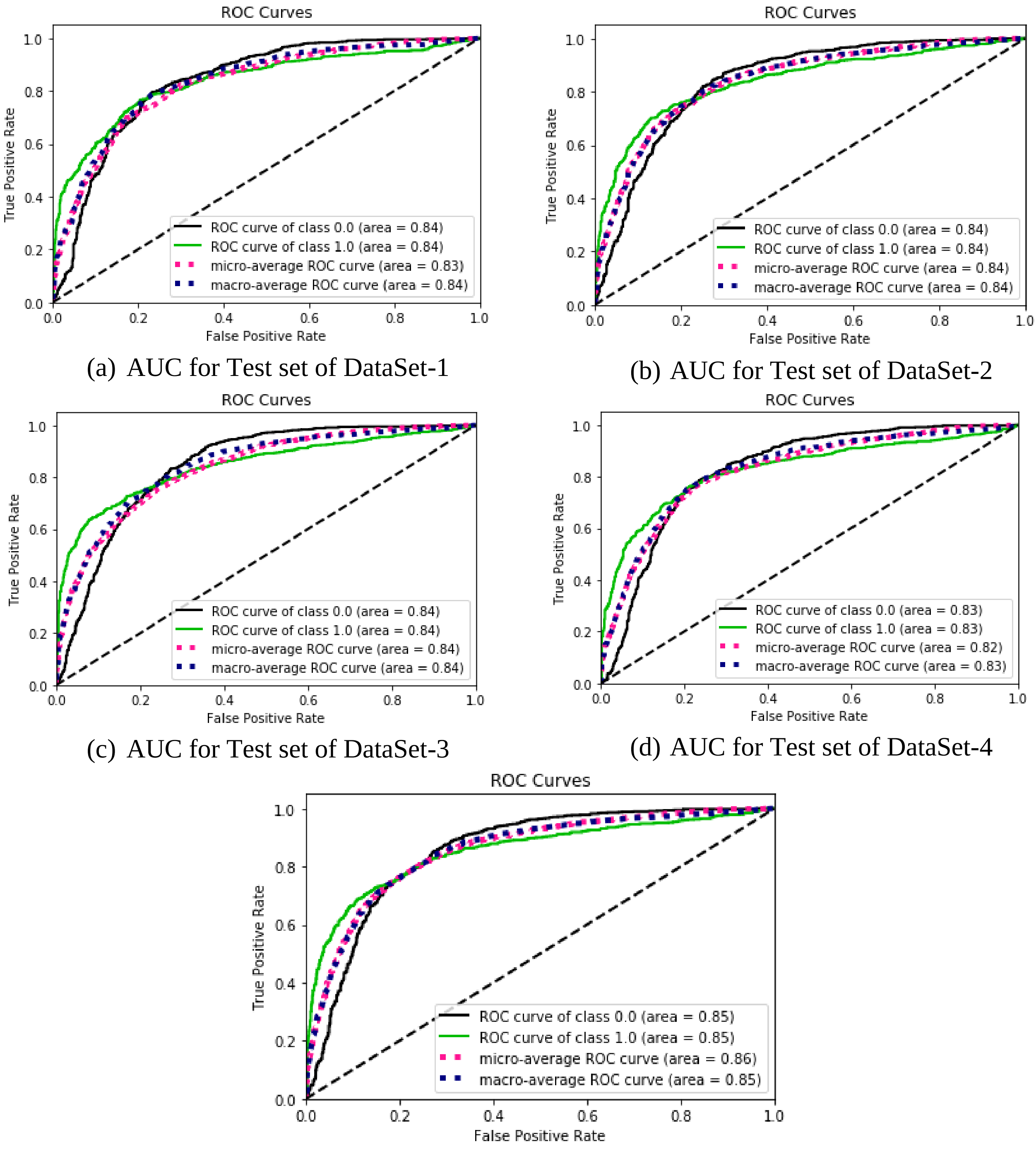
From Tabel.8, Tabel.9, Tabel.10, Tabel.11and Tabel.12we can be seen that in the case of linear kernel the maximum AUC for validation sets is achieved when the nu is **0.5** in all the Datasets. So, nu=0.6 is been chosen as optimal value for the linear kernel.

Now using the **RBF kernel and nu=0.5** the NuSVM is been trained on DataSet-1, DataSet-2, DataSet-3, DataSet-4, and DataSet-5 and AUC was AUC was achieved using Test sets of the Datasets.

Tabel.13: **Tabel.7**: RBF Kernel using Testing

Test No	1	2	3		4	5
AUC	0.84	0.84	0.84		0.83	0.85
Average AUC	0.84		Standard Deviation	0.00632		

Table. 13 shown the AUC on the test sets. The average AUC is **0.84** and the standard deviation is 0.00632. Fig.2 shown the AUC of the test sets.



(a) AUC for Test set of DataSet-1

(b) AUC for Test set of DataSet-2

(c) AUC for Test set of DataSet-3

(d) AUC for Test set of DataSet-4

(e) AUC for Test set of DataSet-5

Fig.2: AUC for the test sets using RBF kernel.

❖ Case: Polynomial Degree=2 Kernel

Different values of nu were used with different datasets that we have shown in the earlier dataset section. Each of the dataset is divided into 40% training and validation set. 40% training data was being used to train the NuSVM and 40% validation was being used to finding the AUC. The results shown in Tabel.14, Tabel.15, Tabel.16, Tabel.17and Tabel.18.

Tabel.14: Linear Kernel using Training: DateSet-1

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.29	0.76	0.81	0.83	0.80	0.77	0.76

Tabel.15: Linear Kernel using Training: DateSet-2

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.22	0.35	0.80	0.84	0.81	0.77	0.76

Tabel.16: Linear Kernel using Training: DateSet-3

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.36	0.68	0.79	0.84	0.80	0.77	0.76

Tabel.17: Linear Kernel using Training: DateSet-4

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.35	0.31	0.82	0.83	0.80	0.77	0.76

Tabel.18: Linear Kernel using Training: DateSet-5

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.61	0.75	0.79	0.84	0.79	0.77	0.76

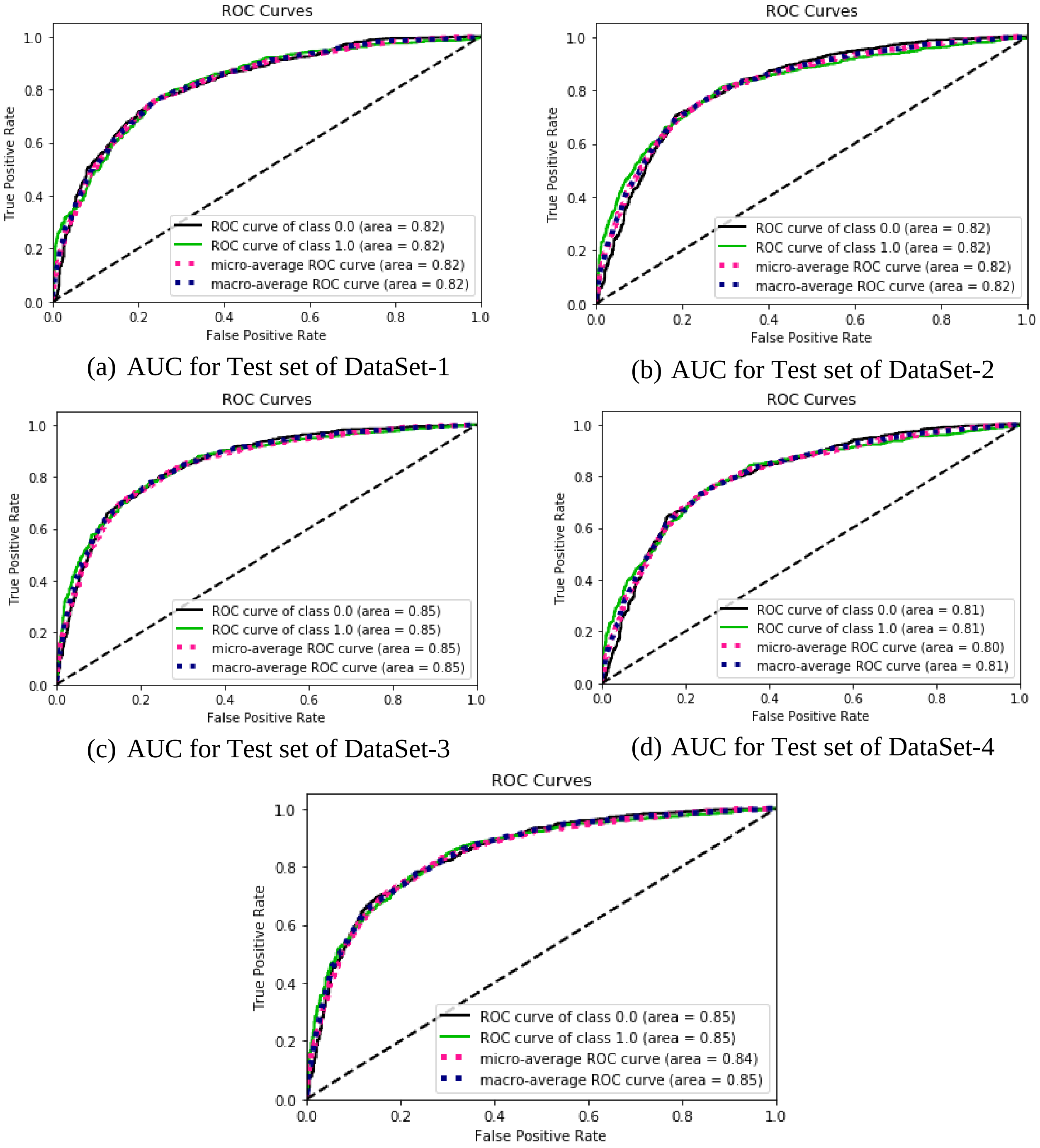
From Tabel.8, Tabel.9, Tabel.10, Tabel.11and Tabel.12we can be seen that in the case of Polynomial of degree 2 kernel the maximum AUC for validation sets is achieved when the nu is **0.6** in all the Datasets. So, nu=0.6 is been chosen as optimal value for the linear kernel.

Now using the polynomial of degree 2, kernel and nu=0.6 the NuSVM is been trained on DataSet-1, DataSet-2, DataSet-3, DataSet-4, and DataSet-5 and AUC was AUC was achieved using Test sets of the Datasets.

Tabel.19: Polynomial of degree 2 Kernel using Testing

Test No	1	2	3	4	5
AUC	0.82	0.82	0.85	0.81	0.85
Average AUC	**0.83**		**Standard Deviation**	**0.0167**	

Table. 19 shown the AUC on the test sets. The average AUC is **0.83** and the standard deviation is **0.0167**. Fig.2 shown the AUC of the test sets.



(a) AUC for Test set of DataSet-1

(b) AUC for Test set of DataSet-2

(c) AUC for Test set of DataSet-3

(d) AUC for Test set of DataSet-4

(e) AUC for Test set of DataSet-5

Fig.3: AUC for the test sets using Polynomial of degree 2 kernel.

❖ Case: Sigmoid Kernel

Different values of nu were used with different datasets that we have shown in the earlier dataset section. Each of the dataset is divided into 40% training and validation set. 40% training data was being used to train the NuSVM and 40% validation was being used to finding the AUC. The results shown in Tabel.20, Tabel.21, Tabel.22, Tabel.23 and Tabel.24

Tabel.20: Linear Kernel using Training: DateSet-1

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.46	0.56	0.63	0.68	0.77	0.77	0.76

Tabel.21: Linear Kernel using Training: DateSet-2

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.48	0.58	0.64	0.70	0.78	0.77	0.76

Tabel.22: Linear Kernel using Training: DateSet-3

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.50	0.57	0.63	0.70	0.78	0.77	0.76

Tabel.23: Linear Kernel using Training: DateSet-4

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.50	0.58	0.65	0.70	0.78	0.77	0.76

Tabel.24: Linear Kernel using Training: DateSet-5

nu	0.1	0.3	0.5	0.6	0.7	0.8	0.9
AUC	0.52	0.59	0.65	0.70	0.78	0.76	0.76

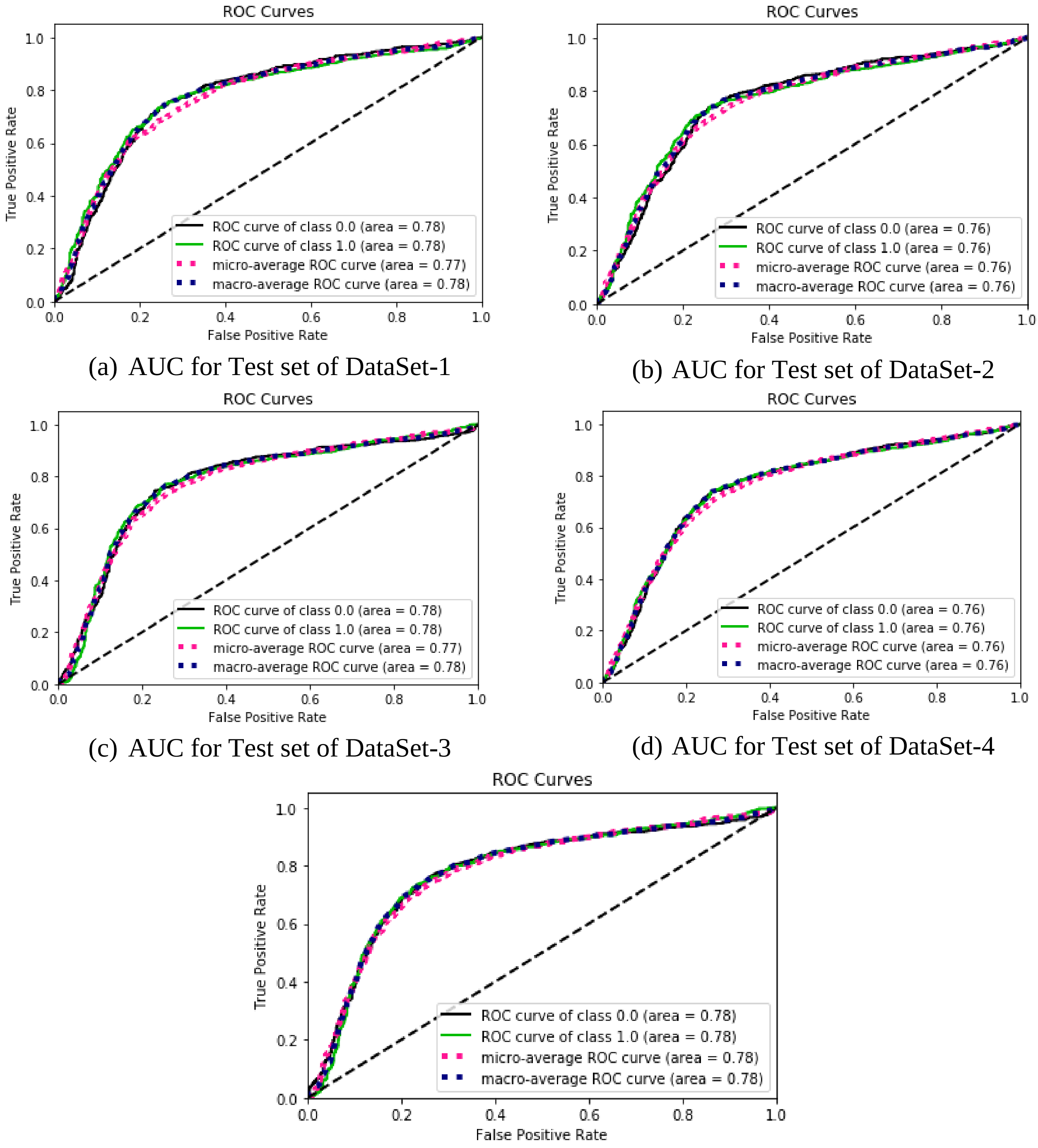
From Tabel.8, Tabel.9, Tabel.10, Tabel.11and Tabel.12we can be seen that in the case of sigmoid kernel the maximum AUC for validation sets is achieved when the nu is **0.6** in all the Datasets. So, nu=0.7 is been chosen as optimal value for the linear kernel.

Now using the **sigmoid kernel** and **nu=0.7** the NuSVM is been trained on DataSet-1, DataSet-2, DataSet-3, DataSet-4, and DataSet-5 and AUC was AUC was achieved using Test sets of the Datasets.

Tabel.25: Sigmoid Kernel using Testing

Test No	1	2	3	4	5
AUC	0.78	0.76	0.78	0.76	0.78
Average AUC	**0.772**		**Standard Deviation**		**0.00979**

Table. 25 shown the AUC on the test sets. The average AUC is **0.772** and the standard deviation is **0. 00979**. Fig.2 shown the AUC of the test sets.



(a) AUC for Test set of DataSet-1

(b) AUC for Test set of DataSet-2

(c) AUC for Test set of DataSet-3

(d) AUC for Test set of DataSet-4

(e) AUC for Test set of DataSet-5

Fig.4: AUC for the test sets using Sigmoid kernel.

Discussion:

Comparing All the cases of using different kernel with the optimal value of nu of the above experiments we can see the nu=0.5 with RBF kernel gave the height Average AUC 84% and the second height average AUC was obtained 83% using nu=0.6 with polynomial kernel with degree 2 both have a standard deviation of 0.0063.From the standard deviation it can be said that the AUC are quite stable in the term of outcomes. Average AUC of using nu=0.6 with linear kernel is 80% and lowest average is 77% which is given by nu=0.7 with sigmoid kernel.

Problem 3.2 [40%]: Use the dataset of Problem 3.1 and perform several splits into a training set and a test set (also with different sizes) to determine which combination of options among the ones you considered in Problem 3.1 ensures the highest AUC. Once you picked out the best model, save (see note below) and submit it together with your report. Your model will be run on a separate matrix containing new test data. Your grade will be based on the performance of your classifier on the new test data, which will contain a very large number of examples generated from the same distribution.

Answer: Among the Above tested NuSVM models, The Model with the nu=0.5 and RBF kernel providing the best results. So, For the best model it has been chosen. Different split was performed on that model with Train and testing data. Train 85% and test 15% providing the best accuracy of around 86%.

This model is being saved. Please find the saved file in the folder upload.