Information Technology Course
Module Software Engineering
by Damir Dobric / Andreas Pech

ML19/20-3.13: Validate and improve Tests of Existing Algorithms: Logistic Regression Algorithm & Self Organizing Map

Hafiz Maaz Ahmed
hafiz.ahmed@stud.fra-usa.de

Abdul Saboor
saboorabdul3333@gmail.com

Abstract—

Machine learning algorithms, namely, Logistic Regression and Self-Organizing Map were used on the previously available and new experimental datasets to validate the existing algorithms. For Logistic Regression we created new tests such as Breast Cancer Diagnosis and Social Network Ads leading to Purchase, to model a logistic regression function (system) using the existing algorithm to produce positive test cases. For Self-Organizing map new datasets of more than two dimensions were employed. The experiment was done through including new UnitTests using MSTest testing platform of Microsoft Dot Net Core Framework. From the successful implementation of tests, we verified the efficiency of existing Logistic regression and Self-Organizing Map algorithms.

Keywords—Logistic Regression, Self-Organizing Map, Machine Learning, Unit Test, MSTest Framework

I. INTRODUCTION

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model further several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

Logistic regression is better than traditional Ordinary Least Squares (OLS) or Linear Function Discriminant Function Analysis for handling dichotomous (two branch outcomes) because it does not rely on strict statistical assumptions such as linearity, continuity, normality or multivariate normality. Furthermore, with the increasing processing speed of the computers, development of sophisticated software to handle complex statistical data, the use of logistic regression is increasing in data analysis.[1]

A self-organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

1.1 Logistic Regression Theory

The central mathematical concept that underlies logistic regression is the logit—the natural logarithm of an odds ratio. The simplest example of a logit derives from a 2 × 2 contingency table. Consider an instance in which the distribution of a dichotomous outcome variable (a child from an inner-city school who is recommended for remedial reading classes) is paired with a dichotomous predictor variable(gender). Example data are included in Table 1. A test of independence using chi-square could be applied. The results yield $\chi2(1) = 3.43$. Alternatively, one might prefer to assess a boy's odds of being recommended for remedial reading instruction relative to a girl's odds. The result is an odds ratio of 2.33, which suggests that boys are 2.33 times more likely, than not, to be recommended for remedial reading classes compared with girls. The odds ratio is derived from two odds (73/23 for boys and 15/11 for girls); its natural logarithm [i.e., ln(2.33)] is a logit, which equals 0.85. The value of 0.85 would be the regression coefficient of the gender predictor if logistic regression were used to model the two out-comes of a remedial recommendation as it relates to gender.

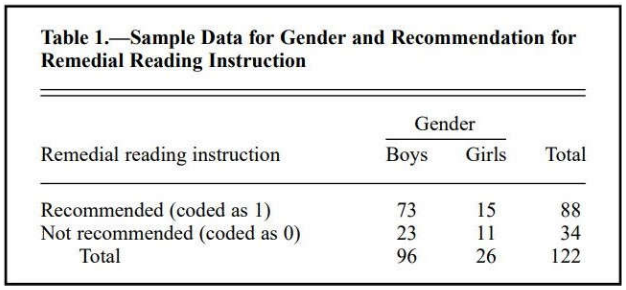
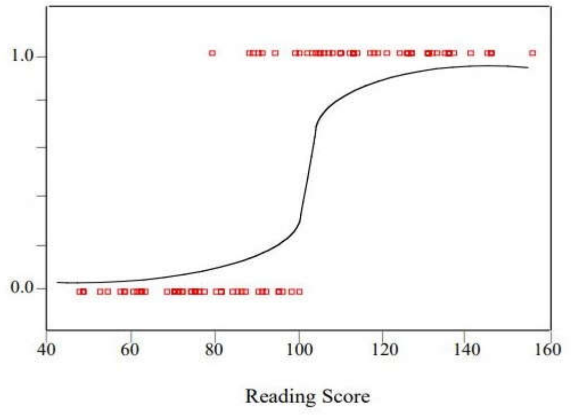


Table 1.—Sample Data for Gender and Recommendation for Remedial Reading Instruction

	Gender		
Remedial reading instruction	Boys	Girls	Total
Recommended (coded as 1)	73	15	88
Not recommended (coded as 0)	23	11	34
Total	96	26	122

Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. In the simplest case of linear regression for one continuous predictor X (a child's reading score on a standardized test) and one dichotomous outcome variable Y (the child being recommended for remedial reading classes), the plot of such data results in two parallel lines, each corresponding to a value of the dichotomous outcome(Figure 1).[2]

Figure 1. Relationship of a Dichotomous Outcome Variable, *Y* (1 = Remedial Reading Recommended, 0 = Remedial Reading Not Recommended) With a Continuous Predictor, Reading Scores



1.2 Self-Organization Map Theory

The Self-Organizing Map (SOM) represents a distribution of input data items using a finite set of models. In the SOM, these models are automatically associated with the nodes of a regular (usually two-dimensional) grid in an orderly fashion such that more similar models become automatically associated with nodes that are adjacent in the grid, whereas less similar models are situated farther away from each other in the grid. This organization, a kind of similarity diagram of the models, makes it possible to obtain an insight into the topographic relationships of data, especially of high-dimensional data items. If the data items belong to certain predetermined classes, the models (and the nodes) can be calibrated according to these classes. An unknown input item is then classified according to that node, the model of which is most similar with it in some metric used in the construction of the SOM.[3]

II. LOGISTIC REGRESSION TESTS

2.1 METHODOLOGY

For testing this Machine Learning Algorithm (Logistic Regression Experiment), two distinct data samples were added to the LearningApi project, namely,

a) Social Network Ads leading to Purchase of item
b) Breast Cancer Diagnosis

Experiment: Social Network Ads leading to Purchase of item

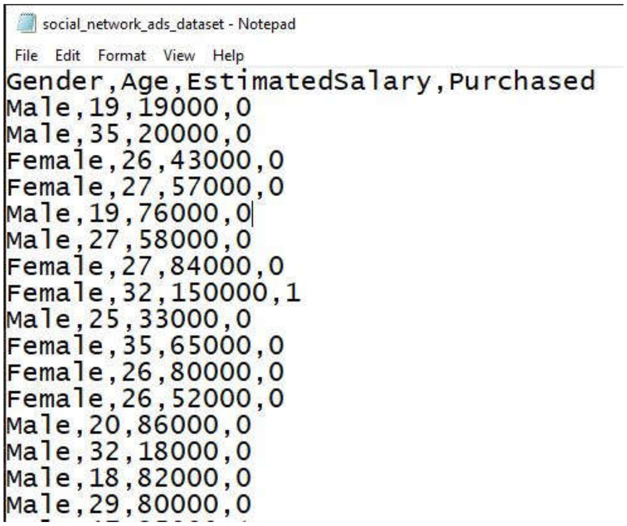
This dataset shows whether a person after watching an ad on Social Network will make a purchase or not.

In this dataset, we have three independent variables referring to a person i.e.

i) Gender (binary value: male or female)
ii) Age (numeric value)
iii) Estimated Salary (numeric value)

Which has dichotomous (two outcome) whether the person will make a purchase or not.

The dataset file social_network_ads_dataset.csv has been added to LearningApi project following the path: LearningApi\LearningApi\test\LearningApiTests\bin\Debug\netcoreapp2.2\SampleData\binary\social_network...csv



```
Gender,Age,EstimatedSalary,Purchased
Male,19,19000,0
Male,35,20000,0
Female,26,43000,0
Female,27,57000,0
Male,19,76000,0
Male,27,58000,0
Female,27,84000,0
Female,32,150000,1
Male,25,33000,0
Female,35,65000,0
Female,26,80000,0
Female,26,52000,0
Male,20,86000,0
Male,32,18000,0
Male,18,82000,0
Male,29,80000,0
```

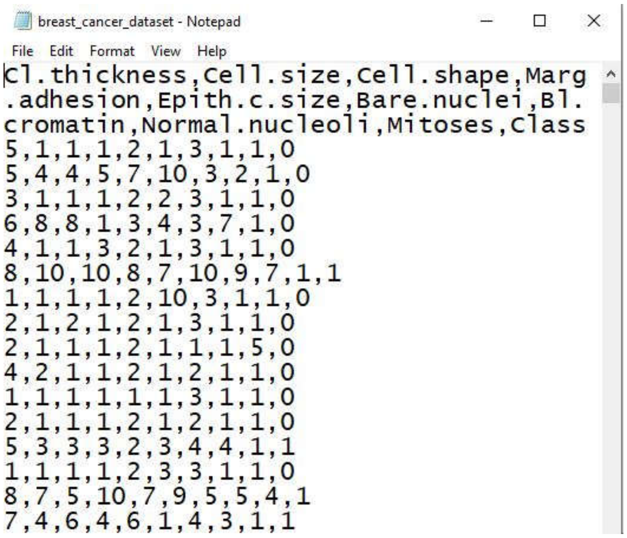
Fig. 2.1: Data sample for Social Networks Ads Experiment.

Experiment: Breast Cancer Diagnosis

This dataset shows if a person has malignant (or benign) cells and diagnosed with having cancer diseases after examining different characteristics of the abnormal cells.

In this dataset, we have nine independent variables (numeric values) referring to the cell characteristics i.e. Cell thickness, Cell size, Cell shape, Marginal Adhesion etc.

The dataset file breast_cancer_dataset.csv has been added to LearningApi project following the path: LearningApi\LearningApi\test\LearningApiTests\bin\Debug\netcoreapp2.2\SampleData\binary\ breast_cancer_dataset.csv



```
Cl.thickness,Cell.size,Cell.shape,Marg
.adhesion,Epith.c.size,Bare.nuclei,Bl.
cromatin,Normal.nucleoli,Mitoses,Class
5,1,1,1,2,1,3,1,1,0
5,4,4,5,7,10,3,2,1,0
3,1,1,1,2,2,3,1,1,0
6,8,8,1,3,4,3,7,1,0
4,1,1,3,2,1,3,1,1,0
8,10,10,8,7,10,9,7,1,1
1,1,1,1,2,10,3,1,1,0
2,1,2,1,2,1,3,1,1,0
2,1,1,1,2,1,1,1,5,0
4,2,1,1,2,1,2,1,1,0
1,1,1,1,1,1,3,1,1,0
2,1,1,1,2,1,2,1,1,0
5,3,3,3,2,3,4,4,1,1
1,1,1,1,2,3,3,1,1,0
8,7,5,10,7,9,5,5,4,1
7,4,6,4,6,1,4,3,1,1
```

Fig. 2.2: Data sample for Breast Cancer Diagnosis Experiment.

The Logistic Regression was performed on the dataset to train a model. First 50 rows (data samples) were not included in the training model, so they can later be used to implement and verify the resultant Logistic Regression model.

2.2 RESULT

Sample data, excluded from the training model was used to test the Logistic Regression model, as shown in the below code snippet:

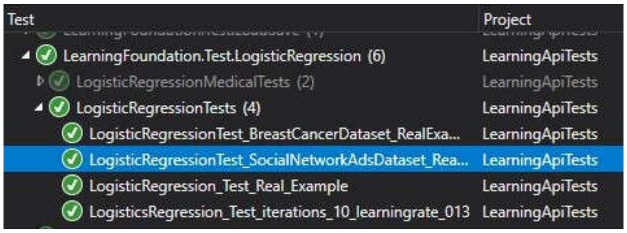
```
apiPrediction.UseActionModule<object[][],
object[][]>((input, ctx) =>
        {
          var data = new object[20][]
          {
            new object[]{"Male", 19, 19000, 0 },
            new object[]{"Male", 35, 20000, 0 },
                .
                .
            new object[]{"Female", 26, 43000, 0 },
          };
          return data;
        });
```

After adjusting the key variables like learning rate and iterations, we used the trained model to predict the correct outputs, as shown in the below snippet:

```
        var result = api.Algorithm.Predict(testData as
double[][], api.Context) as LogisticRegressionResult;

Assert.AreEqual(Math.Round(result.PredictedValues[0]
, 0), 0);
Assert.AreEqual(Math.Round(result.PredictedValues[1]
, 0), 0);
.
.
Assert.AreEqual(Math.Round(result.PredictedValues[1
9], 0), 1);
```

As it can be seen in the passing unit tests:



```
◢ ✓ LearningFoundation.Test.LogisticRegression (6)                 LearningApiTests
  ▷ ✓ LogisticRegressionMedicalTests (2)                           LearningApiTests
  ◢ ✓ LogisticRegressionTests (4)                                  LearningApiTests
      ✓ LogisticRegressionTest_BreastCancerDataset_RealExa...      LearningApiTests
      ✓ LogisticRegressionTest_SocialNetworkAdsDataset_Rea...      LearningApiTests
      ✓ LogisticRegression_Test_Real_Example                       LearningApiTests
      ✓ LogisticsRegression_Test_iterations_10_learningrate_013    LearningApiTests
```

Fig. 2.3 Showing the unit test results for the Logistic Regression Algorithm test class.

III. SELF-ORGANIZATION MAP TESTS

3.1 Methodology

New sets of experiments were included to test the existing Self-Organizing Map Algorithm. The distinct characteristic of these datasets is their multi-dimensionality i.e. 3D, 4D or more dimensional data. Those which were included in the test are:

a) Three-Dimensional RGB color data
b) Four-dimensional data for flower species identification.

Here, each dimension describes a certain characteristic of the data.

Self-Organizing Map maps those multi-dimensional data into 2D plane, where each distinct element can be identified in a certain cluster formation.

The new datasets have been added along with their respective Test methods in LearningApi project. They datasets can be found following the path: \LearningApi\LearningApi\test\LearningApiTests\SelfOrganizingMap\

Experiment: 3D RGB color mapping on 2D plane

In this experiment, a 3D dataset describing various RGB values (0-255) of different colors were used. The data consists of distinct RGB values which can be mapped on 3D plane. The RGB dataset was created in csv file using the approximate color ranges of 7 colors (Red, Orange, Yellow, Green, Grey, Purple and Blue) and introduced into the Self-Organizing Map algorithm.

Experiment: 4D features of Flower mapping on 2D plane

A dataset providing information of four distinct features of three flowers was used in order to the test the Self-Organization Map algorithm. This data was then entered into Self-Organizing model, so it can be mapped on 2D place.

3.2 RESULT

Experiment: 3D RGB color mapping on 2D plane:

As shown in figure 2.3, colors are mapped on 3D RGB planes.

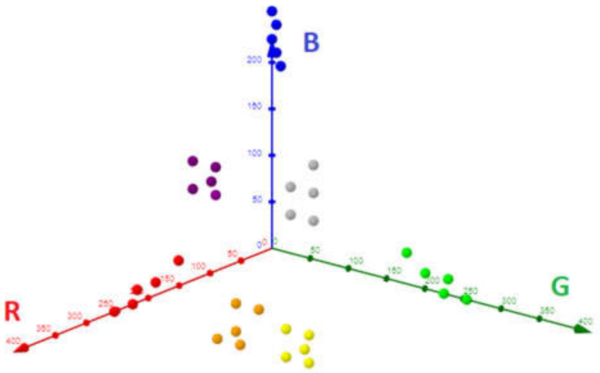


Figure 3.1: Different Colors mapped on RGB (3D) plane

After Self-Organization mapping, the data was mapped on 2D plane, as shown in figure 2.4 below:

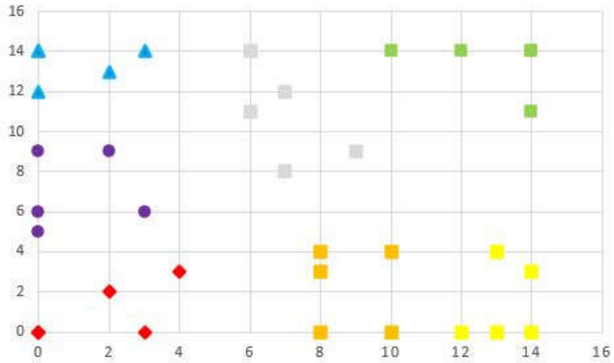
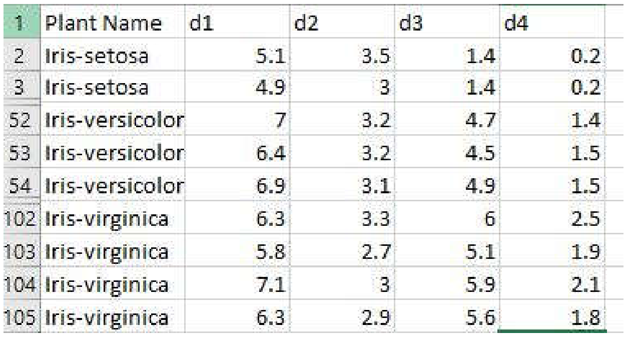


Figure 3.2: Different Color mapped on 2D plane after being entered into Self-Organization Map

From the clustering of the relevant color together, we can visually observe and validate that the machine algorithm of Self Organizing Map.

Experiment: 4D features of Flower mapping on 2D plane

In this experiment, three flower species with 4 distinct features were used as the input dataset to Self-Organizing model, as shown below:



1	Plant Name	d1	d2	d3	d4
2	Iris-setosa	5.1	3.5	1.4	0.2
3	Iris-setosa	4.9	3	1.4	0.2
52	Iris-versicolor	7	3.2	4.7	1.4
53	Iris-versicolor	6.4	3.2	4.5	1.5
54	Iris-versicolor	6.9	3.1	4.9	1.5
102	Iris-virginica	6.3	3.3	6	2.5
103	Iris-virginica	5.8	2.7	5.1	1.9
104	Iris-virginica	7.1	3	5.9	2.1
105	Iris-virginica	6.3	2.9	5.6	1.8

Figure 3.3: Dataset with 4 features (dimensions) of Flower Species

After inputting the dataset to Self-Organization model, following result was obtained, where we can see three flower species each taking respective positions on 2D plane.

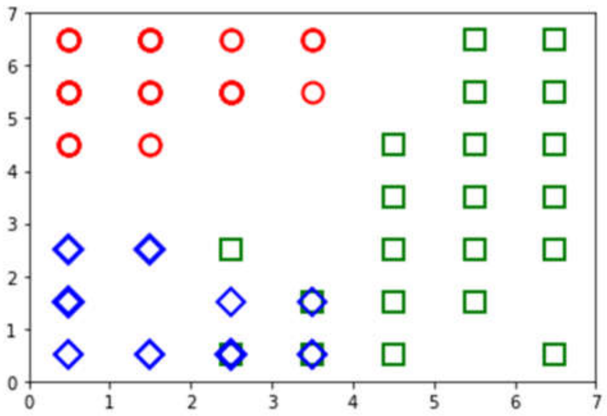


Figure 3.4: Three Flower Species mapped on 2D using 4 distinct features

From the output results given in figure 3.2 and 3.4 we can conclude that the algorithms are working as they are supposed to.

IV. VALIDATING/IMPROVING EXISTING TESTS

New Test Methods were added to perform the new experiments. Some changes were made to the code structure in the test classes to accommodate the sample data as per the requirement. Typing errors were found in the private methods of Test class. For example: same index assignment in the private test method:

```
des.Features[0] = new Column { Id = 1,… };
des.Features[0] = new Column { Id = 2,….};
des.Features[0] = new Column { Id = 3,….};
```

the index [0] of which would result in Test failure for trained algorithm, was corrected:

```
des.Features[0] = new Column { Id = 1,… };
des.Features[1] = new Column { Id = 2,….};
des.Features[2] = new Column { Id = 3,….};
```

New code was added to export the test results of Self-Organization map. So, we can easily plot and visualize test results as it is shown in figure 3.2 and 3.4.

V. CONCLUSION

Logistic Regression and Self-Organization Map Algorithms were found to be working properly, as they were supposed to. New datasets were added to the existing LearningApi project. Using those datasets, new models were trained using the existing machine learning algorithms and prediction from those models were tested via unit testing using MSTest testing platform of Microsoft Dot Net Core Framework.

REFERENCES

[1] Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. Higher Education: Handbook of Theory and Research, Vol. 10, 225–256.

[2] Peng, Lee, Ingersolla Gary M (2002). An Introduction to Logistic Regression Analysis and Reporting in The Journal of Educational Research September 2002

[3] Kohonen, Teuvo (2013). In Twenty-fifth Anniversary Commemorative Issue, Neural Networks January 2013 37:52-65