Assignment-1

Venkata Satya Sai Ajay Daliparthi
veda18@student.bth.se
970316-2637

I. Introduction

The document discusses about implementation of Concept learning through the Least General Generalization (LGG) (algorithms 4.1 and 4.3) for detecting spam mails. Spam base dataset from UCI Machine learning repository is used for this purpose.

II. Methodology

A. Selecting test and train data

We will divide our data set into two parts. Train data and test data.for this purpose we are used sklearn.
We divided data into 80% for training our hypothesis and 20% for testing.
By using KBinsDiscretzer we discritize the data entire data with parameter n_bins = 7.

B. Importing the data set and exploring

By using pandas we imported the data file .then by performing .head() method and .shape on the dataset it is found that there are 4601 rows and 58 columns in our dataset.we performed data.isnull().sum() to check if there are any missing values in our dataset and found out that our data set doesn't contain any missing values. By using KBinsDiscretzer we discritize the data entire data with parameter n_bins = 7.

C. Selecting test and train data

As there are 58 columns in the data set in which 58 of them are instances for a particular sample. We crated a empty list that can store 57 sub lists in it.

D. Training and predicting the results

We calculating hypothesis space:

As each attribute can take seven values

Possible instances are 7^57

Conjugate concept are 8^57

Hypothesis space are 2^7^57

we applied the disjunction inside each conjunction(attribute – sublist of hypothesis). In this way we implemented algorithm 4.3 with 4.1 in main literature.
So Formulated hypothesis is
[0.0, 1.0, 3.0, 2.0, 6.0],[0.0, 1.0, 2.0],[0.0, 1.0, 2.0, 3.0, 5.0, 4.0],[0.0, 1.0, 6.0, 2.0, 5.0, 3.0],[2.0, 0.0, 1.0, 3.0, 4.0, 5.0], [0.0, 1.0, 2.0],[0.0, 1.0, 2.0, 6.0, 5.0, 4.0, 3.0],[0.0, 1.0, 2.0, 6.0],[0.0, 1.0, 2.0, 3.0],[0.0, 1.0, 2.0],[0.0, 1.0, 3.0, 2.0, 4.0, 5.0, 6.0],[0.0, 1.0, 3.0, 2.0, 4.0],[0.0, 1.0, 6.0, 2.0, 3.0], [0.0, 1.0, 2.0, 3.0],[0.0, 3.0, 1.0, 2.0, 6.0],[1.0, 0.0, 2.0, 3.0, 5.0],[0.0, 1.0, 3.0, 5.0, 2.0, 4.0, 6.0],[0.0, 1.0, 3.0, 2.0, 4.0, 6.0],[0.0, 1.0, 2.0, 3.0, 4.0],[0.0, 1.0, 2.0, 6.0],[0.0, 2.0, 1.0, 3.0, 4.0, 6.0],[0.0, 5.0, 1.0, 3.0, 2.0, 6.0, 4.0],[0.0, 4.0, 1.0, 3.0, 2.0, 5.0, 6.0],[0.0, 1.0, 2.0, 3.0, 4.0, 6.0],[0.0, 1.0],[0.0],[0.0, 1.0],[0.0, 1.0, 3.0],[0.0],[0.0, 1.0] [0.0, 1.0, 3.0, 2.0] [0.0], [0.0], [0.0, 1.0],[0.0], [0.0, 1.0],[0.0],[0.0, 1.0],[0.0],[0.0, 1.0],[0.0, 6.0, 3.0],[0.0, 2.0],[0.0, 1.0], [0.0, 1.0, 3.0, 6.0, 4.0, 2.0],[0.0, 4.0, 6.0, 1.0],[0.0, 3.0, 1.0, 2.0, 4.0, 6.0],[0.0, 6.0, 1.0], [0.0, 1.0, 4.0, 6.0]]

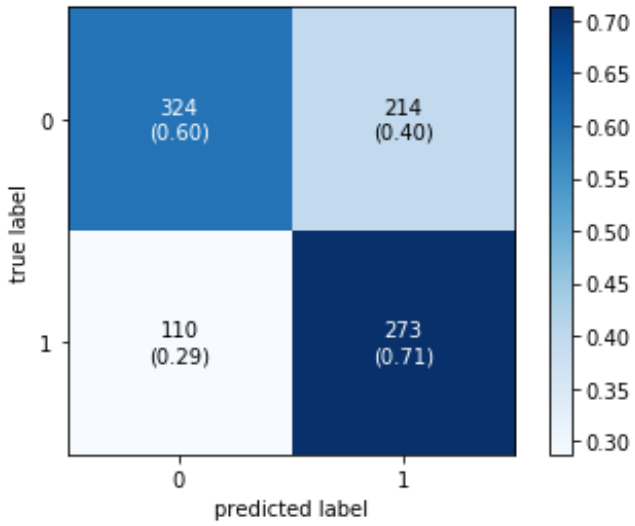
Then after building the hypothesis list we will check all the test values with them and classify weather the sample is spam or ham.

III. Results

We created a list named pred and stored all the predicted values in that list. We have correct values as testing_data_output_list .

To evaluate our model we generated a confusion matrix

Accuracy for the model is ranging between 60 to 65. For example I got accuracy as 64.82 with tp=273, tn=324, fp=214, fn=324



IV. References

[1] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge

University Press, 2012.