Machine learning assignment 1

Harshal Patel hapt18@student.bth.se

Department of computer science

Blekinge institute of technology,

Karlskrona, Sweden

INTRODUCTION

This report aims to present the implementation of concept learner, to categorize mails into spam and not spam. Generally, a concept learner is based on logical models which is sub divided into 2 types namely tree models and role models [1]. Further, the report is divided into sections to explain the details.

DATA

Spambase dataset used in the hypothesis consists of 4600 Instances and 58 attributes. This data is then divided into 80% training and 20% testing. Features from 0 to 57 contains the information about the mail and 58th feature is the class to which it belongs (0-not spam, 1-spam).

METHOD

Data is first converted into data frames using python library. These data frames contain features and class labels separately. Data is then divided into training and testing (80% and 20% respectively). Looping through the testing data, all unique features for a spam (1 - class) mail are acquired. To break the continuity of data, features are divided into 5 bins. Finally, 3680 examples are used for training and 920 examples for testing the hypothesis to find the accuracy.

According to the algorithms 4.1 (least general generalization), a larger set of instance is used for the hypothesis space (set of possible concepts) to make the data more generalized and 4.3 (conjunctive concepts), mentioned in literature [1], explains the conjunction and disjunction, that can be used to reduce redundancy and gather unique features to a class respectively. Resulting hypothesis space is wider, which improves the accuracy when tested on test data.

Finally, accuracy is calculated by counting the number of correctly classified samples and dividing it to total number of test samples.

True positive: correctly classified positive sample.

True negative: correctly classified negative sample.

False positive: incorrectly classified positive sample.

False negative: incorrectly classified negative sample.

Accuracy=(true_positive+true_negative)/(true_positive+true_negative+false_positive+false_negative) *100

Size of hypothesis in the project given we use 5 bins is: $2^{5^{57}}$

RESULTS

Hypothesis is tested using testing data, i.e. 920 samples and accuracy is calculated.

```
number of testing samples: 920
accuracy percentage
62.06521739130435
```

Two-dimensional array formed, represents hypothesis space with all unique spam features of each element is shown below. H1= [0.0 or 1.0 or 2.0 or 4.0] and so on for all features.

```
[[0.0, 1.0, 2.0, 4.0] AND
 [0.0, 1.0] AND
 [0.0, 1.0, 2.0, 3.0] AND
 [0.0, 2.0, 1.0, 4.0] AND
 [0.0, 1.0, 2.0, 3.0] AND
 [0.0, 1.0, 2.0, 3.0] AND
 [0.0, 1.0, 2.0, 3.0, 4.0] AND
 [0.0, 1.0, 2.0, 4.0] AND
 [0.0, 1.0, 2.0, 3.0] AND
 [0.0, 1.0, 4.0, 2.0, 3.0] AND
 [3.0, 0.0, 1.0, 2.0, 4.0] AND
 [0.0, 1.0, 3.0, 2.0] AND
 [0.0, 1.0, 2.0, 4.0] AND
 [0.0, 1.0, 2.0] AND
 [0.0, 1.0, 2.0, 4.0, 3.0] AND
 [0.0, 1.0, 4.0, 2.0] AND
 [1.0, 0.0, 2.0, 3.0, 4.0] AND
 [0.0, 1.0, 2.0, 4.0] AND
 [0.0, 1.0, 2.0] AND
 [0.0, 1.0, 4.0] AND
 [0.0, 1.0, 2.0, 4.0] AND
 [0.0, 3.0, 2.0, 1.0, 4.0] AND
 [0.0, 1.0, 2.0, 3.0, 4.0] AND
 [0.0, 3.0, 1.0, 2.0, 4.0] AND
 [0.0] AND
 [0.0] AND
 [0.0] AND
 [0.0, 1.0, 4.0] AND
 [0.0] AND
 [0.0, 1.0, 2.0, 3.0] AND
 [0.0, 1.0] AND
 [0.0] AND
 [0.0] AND
 [0.0] AND
 [0.0] AND
 [0.0, 1.0] AND
 [0.0, 1.0, 4.0, 3.0] AND
 [0.0, 1.0] AND
 [0.0] AND
 [0.0, 1.0] AND
 [0.0] AND
 [0.0] AND
 [0.0, 1.0] AND
 [0.0] AND
 [0.0, 1.0] AND
 [0.0] AND
 [0.0, 1.0] AND
 [0.0] AND
 [0.0, 1.0] AND
 [0.0, 4.0, 2.0] AND
 [0.0, 1.0] AND
 [0.0, 3.0, 1.0, 2.0, 4.0] AND
 [0.0, 1.0, 4.0, 3.0, 2.0] AND
 [0.0, 4.0, 3.0] AND
 [0.0, 1.0, 2.0, 4.0, 3.0] AND
 [0.0, 4.0] AND
 [0.0, 2.0, 1.0, 4.0, 3.0]]
```

CONCLUSION

Algorithms from [1] (4.1 - LGG and 4.3 - internal disjunction) are successfully implemented to classify spam or not spam mail using the dataset provided. The calculated accuracy shows the performance of the algorithm.

REFERENCE

Flach, P., 2012. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press.