Assignment-1 Concept Learner

Dilukshi Charitha Dissanayake
Department of Computer Science
Blekinge Institute of Technology
Karlskrona,Sweden
didi18@student.bth.se

Keywords—Concept Learner, Least General Generalization

I. INTRODUCTION

This assignment discusses the implementation of a Concept Leaner for the Spambase dataset. Least general Generalization Algorithm 4.1 and 4.2 [1]is used for the implementation.

II. METHODOLOGY

Spambase dataset consists 57 features which I have used to build my model. Since there are 57 features, number of possible instances are 2^{57}, which is '144,115,188,075,855,872'. Therefore, size of the Hypothesis Space is $2^{2\cdot 57.}$ By considering the absence of a feature as an additional value, the possible concepts are 3^{27}, which is '1,570,042,899,082,081,611,640,534,563'.

First, I have binarized my data (0,1) based on the mean value of each feature. If the value is greater than or equlat to the mean value I stored the value as 1 and otherwise as 0. Then, I have divided my dataset into two parts, 80% of the data as training data and 20% is for testing data. Then the training data is been grouped as spam and non-spam. Spam email data were considered as the training data and the non-spam data were included in the testing dataset to make use of the entire dataset.

Based on the spam data, positive instances, I have implemented my concept learner. The first row in the data set was considered as the hypothesis space and then I compared the next row with it. Here I called algorithm 4.2 [1] to compare two instances and if they are not equal I stored the value as '?' and if they are equal I stored the same value. My function returns a hypothesis space afer iterating through all the training data.

After the hypothesis space is returned, I used this to check the accuracy of my model. I passed my hypothesis through the testing data and predicted the outcome. Preditcetd data were stored in a list and then I compared with the actual values.

III. RESULTS

After the model is trained, I iterated it through the testing data. If the hypothesis is met in each data point it returned '1' which means that it is a spam, else '0'for non-spam. In order to check the accuracy of the mode, I have compared the return values (spam=1, non-spam=0) with actual values. Accuracy was based on number of correctly identified emails divided by the total number of tested emails.

TABLE I. ACCURACY

Amount of correct data	***Accuracy***
2789	2789/3159=88.29%

IV. CONCLUSION

I believe that the model seems to be accurate according to the dataset with regard to the simplicity of the model. This model results in 88% accrucy which is more than 50%. Binarizng the data based on mean value seems to be a good factor in predicting spam emails.

REFERENCES

[1] P.Flach, Machine Learning:The Art and Science of Algorithms that Make Sence of Data, Cambrige University Press,201