Assignment 1

Victor Arvidsson
DV2599 Maskininlärning
Blekinge Tekniska Högskola
Karlskrona, Sweden
viar19@student.bth.se

I. Data preprocessing

In order to use a concept learner, we need our data to have discrete values. The dataset that was provided in the assignment contained continuous values for all attributes except the target class and the run_length attributes. However, for the run_length attributes, I decided to discretize these further as well, both to simplify the implementation, but also to reduce the hypothesis space for the data. The target attribute was left untouched.

To discretize the data, I applied the K-Means discretization method using 20 bins for the data. This was chosen experimentally by using different methods and number of bins in order to achieve the best result. The result of these tests can be found in table I. Each bin was assigned a label between 0-19.

Bins	Strategy	Accuracy	Precision	Recall	F1-Score
3	K-means	65.4%	53.7%	89.6%	67.1%
7	K-means	72.2%	62.6%	73.5%	67.6%
15	K-means	77.3%	68.9%	77.0%	72.7%
20	**K-means**	**78.8%**	**69.8%**	**81.3%**	**75.1%**
30	K-means	78.1%	69.0%	80.8%	74.4%
15	uniform	69.9%	58.1%	84.8%	68.9%
15	quantile	72.2%	62.1%	76.1%	68.4%
15	**K-means**	**77.3%**	**68.9%**	**77.0%**	**72.7%**

TABLE I
Experiment results for different number of bins and different strategies. Ran with training size of 1%.

II. Instance-, extension- and hypothesis space

The instance space is defined as the possible number of combinations of the attributes. The dataset contains 57 attributes (excluding the target attribute). With discretization into 20 different bins, this gives us an instance space of

$$20^{57} \approx 1.4 \cdot 10^{74}.$$

The extension space is the number of possible subsets of all instances. The number of subsets of a set is calculated by $2^{|s|}$, where $|s|$ is the cardinality of the set. This gives us an extension space of

$$2^{20^{57}}$$

extensions, which is a number that contains approximately $4.3 \cdot 10^{73}$ digits.

Finally, the hypothesis space, i.e. the number of possible conjunctive concepts, is calculated in the same way as the instance space, but allowing for null values. This gives us a hypothesis space of

$$21^{57} \approx 2.3 \cdot 10^{75}$$

instances.

III. Algorithm and training

For the concept learner, I applied the Least General Generalization (LGG) algorithm on instances classified as spam. I chose to implement the LGG-conj algorithm from the course book (Algorithm 4.2), since this was the most straightforward of the two.

For training, the positive samples from the dataset where selected, i.e. those that were classified as spam. From these, 1% of the samples where selected randomly for training. This was done to ensure that the LGG-algorithm did not generalize to much, and therefore lower it's precision. The 1% was chosen experimentally, and the results from the experiments are shown in table II.

Training data amount	Accuracy	Precision	Recall	F1-score
10%	57.7%	48.2%	97.8%	64.5%
1%	**77.3%**	**68.9%**	**77.0%**	**72.7%**
0.5%	76.0%	68.9%	71.1%	70.0%

TABLE II
Experiment results for different training data amounts. Ran with bin size 15 and K-means strategy.

IV. Result

For the result, the accuracy, precision, recall, and F1-score was calculated. For the spam classification, it is important that we do not have a lot of false positives, as it could lead to users turning the filter off. It is therefore important to measure the precision, in addition to the accuracy. The result for the model can be seen in table III. We also plot the confusion matrix for the classified instances, which can be seen in figure 1. The conjunctive rule for the concept learner can be found in table IV.

Accuracy	Precision	Recall	F1-score
78.8%	69.8%	81.3%	75.1%

TABLE III
Results for the model.

References

[1] P. Flach, “Machine Learning: The Art and Science of Algorithms that Make Sense of Data,” Cambridge University Press, pp. 106–110, 2012.

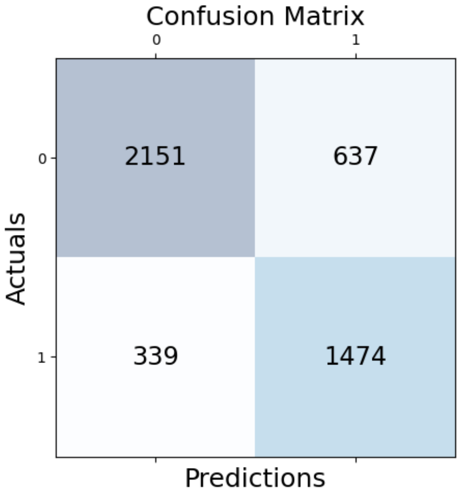

Fig. 1. Confusion matrix for the conceptive learner. 1 is spam and 0 is not spam.

Attribute	**Label**
word_freq_hp	0
word_freq_hpl	0
word_freq_george	0
word_freq_650	0
word_freq_lab	0
word_freq_labs	0
word_freq_telnet	0
word_freq_857	0
word_freq_data	0
word_freq_415	0
word_freq_parts	0
word_freq_pm	0
word_freq_cs	0
word_freq_meeting	0
word_freq_project	0
word_freq_edu	0
word_freq_table	0
word_freq_conference	0
char_freq_[	0

TABLE IV
CONJUNCTIVE RULE FOR THE CONCEPT LEARNER. ALL ATTRIBUTES VALUES ARE PART OF THE CONJUNCTION, AND ALL VALUES NOT PRESENT IN THE TABLE WERE IGNORED.