605.649 Programming Project 1: Winnow-2 and Naïve Bayes

Christopher El-Khouri

1. Introduction

Classification is one of the most common problems that machine learning is used to solve. Many different algorithms have been developed to classify data, each with their own advantages and disadvantages, hence, it is important to know when to use each algorithm effectively.

In this project we will be assessing the performance of two supervised learning algorithms: Winnow-2 and Naïve Bayes. Supervised learning algorithms are algorithms that map an input to an output based on pre-defined input-output pairs.

Winnow-2 is a linear algorithm created by Nick Littlestone (1988) that utilizes weighted coefficients to determine if an array of predictor variables belongs to one of 2 classes. The linear combination of the coefficients with their corresponding variables is calculated and the result is compared with a defined parameter Θ. If the linear combination is greater than Θ then it belongs to one class, otherwise it belongs to the other.

Naïve Bayes is an algorithm that uses Bayes Theorem to classify an array of predictor variables. Bayes Theorem is as follows (Butcher, 2019):

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Where:

$$P(A)\text{: } The\ prior\ probability$$

$$P(A|B)\text{: } The\ posterior\ probability$$

$$P(B|A)\text{: } The\ likelihood$$

$$P(B)\text{: } The\ normalizer$$

In this project we will be comparing the performance of both algorithms on 5 different datasets. We hypothesize that classifiers produced by the Naïve Bayes algorithm will produce more accurate models for all of our different datasets. The accuracy of the models will be calculated by the percentage of correct predictions.

In the next section we will describe the technical and theoretical basis of our algorithms, in section 3 we will present and compare the results, and in section 4 we will discuss the significance of our experiment.

2. Algorithms and Experimental Methods

Winnow-2

The Winnow-2 model is as follows:

$$f(x) = \sum_{i=1}^{d} w_i x_i$$

$$Where:$$

$$w_i: The\ coefficients\ of\ our\ predictor\ variables$$

$$x_i: Our\ predictor\ variables$$

$$Considering\ a\ threshold\ \theta\ where:$$

$$h(x) = \begin{cases} 1, & f(x) > \theta \\ 0, & f(x) \leq \theta \end{cases}$$

When training the model, we would run the algorithm on training data and promote or demote the coefficients by a factor α as follows:

- $if\ f(x) > \theta\ and\ h(x) = 0\ then\ demote$
- $if\ f(x) \leq \theta\ and\ h(x) = 1\ then\ promote$

The promotion and demotion are carried out as described below:

$$Promoting:$$

$$w_i = \begin{cases} \alpha w_i, & x_i = 1 \\ w_i, & x_i = 0 \end{cases}$$

$$Demoting:$$

$$w_i = \begin{cases} w_i/\alpha, & x_i = 1 \\ w_i, & x_i = 0 \end{cases}$$

The α and Θ values are initially chosen by tuning the algorithm on our data set. Tuning is carried out by extracting 10% of our data and training the Winnow-2 algorithm on the data on α values ranging from 1.1 to 10 with increments of 0.1, and on Θ values ranging from 0.5 to 5 with increments of 0.1. The α and Θ values that result in the highest accuracy within the tuning data i.e. the highest amount of correct predictions will be selected for modeling the remaining data.

Naïve Bayes

The Naïve Bayes model is as follows:

$$class = argmax_c \, P(c) \prod_{i=1}^{d} P(f_i|c)$$

$$Where:$$

$$c: class$$

$$d: dimension$$

$$f: feature$$

When training the data, it is possible for some probabilities to result in 0, this may be due that when taking a training set, not all the different combination of variables may be there. Therefore, we can use an approach called smoothing which goes as follows:

- $if \; P(f_i|c) = \frac{n_{f,c}}{n_c} = 0$
- $then, let \; P(f_i|c) = \frac{n_{f,c}+p}{n_c+m}$

The m and p values are initially chosen by tuning the algorithm on our data set. Tuning is carried out by extracting 10% of our data and training the Naïve Bayes algorithm on the data with m values ranging from 1 to 10 with increments of 1, and on p values ranging from 0.001 to 0.01 with increments of 0.001. The m and p values that result in the highest accuracy within the tuning data i.e. the highest amount of correct predictions will be selected for modeling the remaining data.

5-fold cross-validation

Our algorithms were fitted and validated using 5-fold cross validation. After tuning our algorithms using 10% of the data, 90% of the data was used in the 5-fold cross validation process. The data is split into 5 parts (or folds) and the classifiers are trained on each of the folds accordingly. In the case of Winnow-2, the w coefficients resulted from each of the 5 folds considered and the final model’s coefficients is the average of each of the coefficients. Similarly, in Naïve Bayes, the Bayesian probabilities of each of the 5-folds are calculated and the final model’s probabilities is the average.

3. Datasets

The following datasets were used in this project:

1. Breast Cancer

   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
2. Glass

   https://archive.ics.uci.edu/ml/datasets/Glass+Identification
3. Iris

   https://archive.ics.uci.edu/ml/datasets/Iris
4. Soybean (small)

   https://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29
5. Vote

   https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

Dataset 1: Breast Cancer

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The dataset contains 699 points and 2 classes: Benign (denoted by '2') and Malignant (denoted by '4'). Including the class variable 'Class', the data contains a total of 11 features. Furthermore, there are 16 missing values that need pre-processing, they are all under the feature 'Bare Nuclei'. We will fill those missing values by separating the data into the 2 different classes and producing sampled data based on the median of the 'Bare Nuclei' feature of the 2 classes.

The median was chosen due to the distribution of 'Bare Nuclei' data shown below:

Table 1: Statistical Description of Bare Nuclei Data by Class

Class=2	
	bare_nuclei
count	444
mean	1.34684685
std	1.1778482
min	1
25%	1
50%	1
75%	1
max	10

Class=4	
	bare_nuclei
count	239
mean	7.627615063
std	3.116678999
min	1
25%	5
50%	10
75%	10
max	10

Based on the table above, class 2's distribution shows that at least 75% of the data have a value of 1, class 4's distribution shows that at least 50% of the data have a value of 10. The mean values are 1.34 and 7.62 respectively, sampling the data based on the mean values will not give accurate simulations due to the values of 'Bare Nuclei' being round figures and due to the fact that the data is clearly skewed.

The attribute values were discretized by producing boolean columns of whether or not the values are greater or less than the medians.

Dataset 2: Glass

This study of classification of types of glass was motivated by criminological investigation. The dataset contains 214 instances, 6 classes, and 11 columns. The attribute values were discretized by producing boolean columns of whether or not the values are greater or less than the means. The class values were discretized by one-hot encoding the 'Type' column. Since Winnow-2 can only handle 2-class classification, the data was modeled with Winnow-2 one type at a time.

Dataset 3: Iris

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The attribute values were discretized by producing boolean columns of whether or not the values are greater or less than the means. The class values were discretized by one-hot encoding the 'class' column. Since Winnow-2 can only handle 2-class classification, the data was modeled with Winnow-2 one type at a time.

Dataset 4: Soybean (small)

The dataset consists of a small subset of the original soybean database. The dataset contains 47 instances, 4 classes, and 35 columns. The class values were discretized by one-hot encoding the last column. Since Winnow-2 can only handle 2-class classification, the data was modeled with Winnow-2 one type at a time.

Dataset 5: Vote

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac. The dataset contains 435 instances, 2 classes, and 17 columns. The class values were discretized by one-hot encoding the last column. The attribute values were discretized by giving 'y' values a value of 1 and 'n' values a value of 0. The ‘?’ values which represent ‘abstain’ were replaced with the most frequent occurring value of the feature of that respective class i.e. the mode.

4. Results

Dataset 1: Breast Cancer

Modeling the Breast Cancer dataset with Winnow-2 we get the following:

Table 2: Winnow-2 results for Dataset 1

Dataset 1: Breast Cancer	
α	1.1
Θ	0.5
accuracy	75.4%

Modeling the Breast Cancer dataset with Naïve Bayes we get the following:

Table 3: Naive Bayes results for Dataset 1

Dataset 1: Breast Cancer	
m	1
p	0.001
accuracy	96.8%

We see a significant difference in accuracy between the Naïve Bayes classifier and Winnow-2.

Dataset 2: Glass

Modeling the Glass dataset with Winnow-2 we get the following:

Table 4: Winnow-2 results for Dataset 2

Dataset 2: Glass						
Parameter	Type 1	Type 2	Type 3	Type 5	Type 6	Type 7
α	1.1	1.1	1.1	1.1	1.1	1.1
Θ	0.6	0.6	0.5	0.5	0.5	0.5
accuracy (%)	63.2	55.3	92.1	92.1	97.4	94.7

Modeling the Glass dataset with Naïve Bayes we get the following:

Table 5: Naive Bayes results for Dataset 2

Dataset 2: Glass	
m	1
p	0.001
accuracy (%)	63.2

Dataset 3: Iris

Modeling the Iris dataset with Winnow-2 we get the following:

Table 6: Winnow-2 results for Dataset 3

Dataset 3: Iris			
Parameter	Class 1	Class 2	Class 3
α	1.1	1.1	1.1
Θ	0.5	0.5	0.7
accuracy (%)	63.0	66.7	70.4

Modeling the Iris dataset with Naïve Bayes we get the following:

Table 7: Naive Bayes results for Dataset 3

Dataset 3: Iris	
m	1
p	0.001
accuracy (%)	77.8

Dataset 4: Soybean (small)

Modeling the Soybean dataset with Winnow-2 we get the following:

Table 8: Winnow-2 results for Dataset 4

Dataset 4: Soybean				
Parameter	D1	D2	D3	D4
α	1.1	1.1	1.1	1.1
Θ	1	0.5	1	0.5
accuracy (%)	62.5	37.5	75	100

Modeling the Soybean dataset with Naïve Bayes we get the following:

Table 9: Naive Bayes results for Dataset 4

Dataset 4: Soybean	
m	1
p	0.001
accuracy (%)	100

Dataset 5: Vote

Modeling the Vote dataset with Winnow-2 we get the following:

Table 10: Winnow-2 results for Dataset 5

Dataset 5: Vote	
α	1.1
Θ	0.5
accuracy (%)	50

Modeling the Vote dataset with Naïve Bayes we get the following:

Table 11: Naive Bayes results for Dataset 5

Dataset 5: Vote	
m	1
p	0.001
accuracy (%)	88.5

5. Discussion

As shown on the tables above, our hypothesis appears to stand with Naïve Bayes outperforming the Winnow-2 algorithm. Besides the fact that Naïve Bayes is superior in handling data with multiple classes, Winnow-2 seems to return a high accuracy when most of the values that are to be predicted are 0.

However, Winnow-2 did extremely well with some of the Soybean data classes. This could be either due to the large number of columns or due to the discrete values of the features ranging from 0-10. On the other hand, Naïve Bayes did not have an accuracy of less than 60% and even reached 100% in the Soybean dataset.

6. Conclusion

In conclusion, we can safely say that Naïve Bayes is an overall superior classification algorithm to Winnow-2 and consider our hypothesis to be true. Perhaps further research into the optimization of Θ and α values and into discretization methods improve the performance of the algorithm. Furthermore, Winnow-2 only being able to handle 2-class classification problems could prove to be complicated in most practical applications.

7. References

Nick Littlestone.
Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm
Machine Learning 2, 285–318 (1988)

Stephyn Butcher.
Fundamentals of Data Science
Release Spring 2019.1