Contents

Project 1: Comparison Between Winnow 2 and Bayes Algorithm

Jaudat Raza JAUDAT.RAZA@JHU.EDU
Master of Computer Science
John Hopkins University
Baltimore, MD 21218

Editor: Jaudat Raza

Abstract

This document will cover the process of creating Winnow 2 and Bayes Algorithm and run 5 data sets on each algorithm to see how they perform. The resource used will be linked below in the resource section of the document. The document will follow a simple pattern for which a less than 5 video will be submitted as well. The Scripts start with getting the data into Pandas Data from the link provided in the data source. Then it will clean the data, so there aren't any null or dead data point. After that, the data will be processed with Winnow 2 Algorithm and then it will be run through Bayes Algorithm. The Training set and Test Set will use two different process so that will be a variable of the input, but it should not impact the result we expect in our hypothesis that Bayes should perform better then Winnow 2

1 Problem Statement

The purpose of this project was to Learn Winnow 2 and Bayes Algorithm. These are beginning level Algorithm, but Winnow being very simplistic is expected to perform less efficient and general method compare to Bayes. We will run the data from the UCI data set to see the result of which algorithm performs better in the Test. Based on that we will conclude which algorithm is more efficient.

2 Algorithm Studied

In this project we are covering Winnow 2 and Bayes algorithm.

2.1 Winnow 2

Winnow 2 is a supervised learning Algorithm. The goal of the algorithm is to go through each instance of the data set and be able to start predicting better and better as more data is processing thought the algorithm.

The model receive one instance of the data and Winnow 2 only works with 0 or 1, so you should preprocess your data for the algorithm to work better on your data. Then using the previous value, the algorithm makes the prediction for the next set of instances. So if the model makes the right prediction then, the model is left alone, but if it makes the wrong prediction, it could be wrong in two ways.

1) False Negative where Prediction was 0 and the correct answer was 1
   If this is the case, then the step the Model takes is to promote the value
2) False Positive where the prediction was 1 and the correct answer was 0
   If this is the case, then the step the Model takes is to Demote the value

Runs through Weighted Sum :

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

Where:

- d is the total number of attributes
- w_i is the weighting of the i^{th} attribute
- x_i is the value of the i^{th} attribute in binary format
- f(x) is the weighted sum (e.g. $w_1x_1 + w_2x_2 + w_3x_3 + ... + w_dx_d$)

$$h(x) = 1\ if\ f(x) > 0$$

$$h(x) = 0\ if\ f(x) \leq 0$$

- h(x) is the predicted class
- Θ is a constant threshold (commonly set to 0.5)

If the learner makes an incorrect prediction, either promotion or demotion occurs. In both promotion and demotion, the weights w_i are modified by using a constant parameter α which is any value greater than 1, so we are using 2 for our model in the code and for theta we are using 0.5

Then perform this till the end.

2.2 Naïve Bayes Rule

The Naive Bayes algorithm is a technique based on Bayes Theorem for calculating the probability of a hypothesis (H) given some pieces of evidence (E).

Mathematical Bayes Theorem

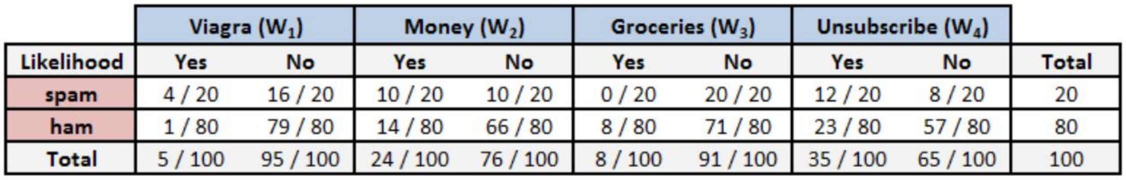
$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

Where:

P(B|A) = posterior probability: the probability of a hypothesis after taking the evidence into account (e.g. probability of being sick given all this evidence)

P(A|B) = likelihood: the probability of the evidence given the hypothesis (e.g. probability of having red eyes given that a person is sick)

P(B) = class prior probability: the known probability of the hypothesis (e.g. probability of being sick for the population or entire sample of instances)



	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		
Likelihood	Yes	No	Yes	No	Yes	No	Yes	No	Total
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

This is common example of how each probability is accounted given the circumstances.

3 Tuning

This is the section that describes all the data and how it was updated to make it work with the Algorithm. This will be described in the video as well.

3.1 Breast Cancer.

1) Dropped the dead data from the set. It was marked in question marked so it was removed in the preprocess section
2) BareNuclei Attribute of the data was converted to numric so readable by the algorithm
3) If Benign the Class was marked as 1
4) If Malignant the Class attribute was marked as 0

Glass

3.2 Glass

1) No Data Change just made sure it was processed correctly.

3.3 Iris

1) If Iris-setosa was in the class the class value was changed to 0
2) If Iris-virginica was in the class the class value was changed to 1
3) If Iris- versicolorwas in the class the class value was changed to 2

3.4 SoyBean

1) Removed columns that did nothing which included
    a. leafspotshalo
    b. leafshread
    c. leafmalf
    d. leafmild
    e. seed
    f. moldgrowth
    g. seeddiscolor
    h. seedsize
    i. shriveling
2) Updated the Distribution to 1, 2, 3, 4

3.5 Vote

1) Dead Data dropped, which was a big loss . In updates it could be replaced by mean or something else to make the data better. 202 entries deleted.
2) One Hot Encoding on Class for demo and Republican
3) Rest of the attributes y is 1 and no was 0

4 Results

4.1 Breast Cancer.

4.1.1 Winnow

```
These are the results from running the algorithm
Number of Test Instances : 225
True Positives : 0
False Positives : 85
False Negatives : 0
True Negatives : 140
Accuracy : 62.22222222222222%
```

4.1.2 Bayes

```
---------------------------------------------------------
Naive Bayes Summary Statistics
---------------------------------------------------------
Data Set : breast_cancer.txt


Accuracy Statistics for All 5 Experiments:
[0.99270073 1.         0.95620438 0.98529412 0.97777778]


Classification Accuracy : 98.23954009827777%
```

4.2 Glass

4.2.1 Winnow

```
These are the results from running the algorithm
Number of Test Instances : 70
True Positives : 21
False Positives : 0
False Negatives : 0
True Negatives : 49
Accuracy : 100.0%
```

4.2.2 Bayes

```
---------------------------------------------------------
Naive Bayes Summary Statistics
---------------------------------------------------------
Data Set : Glass.txt


Accuracy Statistics for All 5 Experiments:
[0.31111111 0.31818182 0.3255814  0.33333333 0.33333333]


Classification Accuracy : 32.43081982616866%
```

4.3 Iris

4.3.1 Winnow

```
These are the results from running the algorithm
Number of Test Instances : 49
True Positives : 14
False Positives : 19
False Negatives : 0
True Negatives : 16
Accuracy : 61.224489795918366%
```

4.3.2 Bayes

```
---------------------------------------------------------
Naive Bayes Summary Statistics
---------------------------------------------------------
Data Set : Iris.txt


Accuracy Statistics for All 5 Experiments:
[0.93333333 0.9        0.76666667 0.63333333 0.79310345]


Classification Accuracy : 80.52873563218391%
```

4.4 Soy Bean

4.4.1 Winnow

```
These are the results from running the algorithm
Number of Test Instances : 15
True Positives : 4
False Positives : 0
False Negatives : 0
True Negatives : 11
Accuracy : 100.0%
```

4.4.2 Bayes

```
---------------------------------------------------------
Naive Bayes Summary Statistics
---------------------------------------------------------
Data Set : SoyBean.txt


Accuracy Statistics for All 5 Experiments:
[1.         0.9        0.88888889 0.88888889 1.        ]


Classification Accuracy : 93.55555555555554%
```

4.5 Vote

4.5.1 Winnow

These are the results from running the algorithm

```
Number of Test Instances : 77
True Positives : 54
False Positives : 1
False Negatives : 10
True Negatives : 12
Accuracy : 85.71428571428571%
```

4.5.2 Bayes

```
---------------------------------------------------------
Naive Bayes Summary Statistics
---------------------------------------------------------
Data Set : Vote.txt


Accuracy Statistics for All 5 Experiments:
[0.72340426 0.74468085 0.78723404 0.82608696 0.77777778]


Classification Accuracy : 77.18367766471374%
```

5 Conclusion

This isn’t as clear as the answer as i expected, but as we can see that usually Bayes has more consistently don’t better the Winnow 2 in most data sets. Each fold for Bayes did improve the accuracy of the result and accuracy of our model prediction capability.

Resources

Joyce, James. “Bayesâ□□ Theorem.” *Stanford Encyclopedia of Philosophy*, Stanford University, 30 Sept. 2003, plato.stanford.edu/entries/bayes-theorem/.