Anthony Gray
EN.605.649.81
Intro to Machine Learning
Project 1

Abstract
For this project, the Naïve Bayes algorithm and a modified variant of the Winnow-2 algorithm were implemented and tested on five public datasets. The Naïve Bayes algorithm performed with consistent accuracy on four of the five sets, while the Winnow-2 algorithm scored poorly on all of them, likely due to an implementation error.

Problem Statement and Algorithms
In this project two classification algorithms are each applied to five different data sets. The algorithms are the Winnow-2 (6) and Naïve Bayes (7) algorithms. Each is an online, supervised learning algorithm that operates by generating a model based on training data that can be used to classify previously unobserved instances.
Winnow-2 works by producing a vector of weights, one per attribute in the target instance, using the sum of the weighted values of the instance to produce a classification, and updating those weights based on the accuracy of that classification. For this exercise, Winnow-2 was modified to produce a different weight vector for each class in the dataset. The weights were generated and used in the normal way except that the algorithm would attempt to classify each instance as the target class or as not the target class with each weight vector.
The Naïve Bayes algorithm works by calculating the probability of each class and the conditional probability of each attribute by each class based on the frequency of the classes and attributes in the training data. It then uses these probabilities to calculate the probability of a given instance belonging to each class based on its attributes and assigns the classification with the greatest probability.
Both algorithms were expected to perform reasonably well, with variation depending on their suitability to the specifics of each data set. Specifically, the Winnow-2 algorithm was expected to perform best when only two classes were present in the data, as it is designed to make a choice between two options. The Naïve Bayes algorithm is likewise designed to make a binary decision, though it does work with more than two classes. It was therefore expected to perform better on those data sets with a greater number of classes.

Procedure/Tuning
Five datasets were used:

Name	**Instances**	**Attributes**	**Classes**
Iris (1)	150	4	3
Glass (2)	214	10	7
House Votes (3)	435	16	2
Breast Cancer (4)	699	10	2
Soybeans (5)	47	35	4

Each dataset was downloaded from its public repository and treated for use by the two algorithms. Missing data in the House and Cancer sets were replaced by the respective column mean. The first columns of the Glass and Cancer sets were dropped, as they reference sample identification numbers and would not contribute to classification. Column order was rearranged where necessary to ensure that the Class value was always in the last column, and the data types were changed to numeric for all values except the Class. Finally, because both algorithms operate over Boolean variables, the data in all datasets except the House set were converted to Booleans by finding the mean of the column and changing each

item to a 1 if it was greater than the mean or a 0 if it was less than the mean. Both algorithms were implemented, and their accuracies compared when applied to the same five data sets.
The sets were split by thirds, with two thirds acting as the training set, and the remaining third serving as the test set. Each algorithm pair was run on each dataset five times and the results recorded in a spreadsheet. The models produced by the learning methods were also recorded in a text file.

Results

	Winnow Averages			Naïve Bayes
Data	**Classification Rate %**	**Accuracy %**	**Total Accuracy %**	**Average Accuracy %**
Iris	49.6	65.8	32.6368	77.4
Glass	69.6	69.8	48.5808	56
House	100	100	100	91.2
Cancer	36	32.2	11.592	97.4
Soybean	100	44.2	44.2	97.2

This table was generated from the results of five runs over each complete dataset, though each run resulted in a different random split between training and test sets. To account for sets with more than two classes, the winnow algorithm attempted to classify each class in sequence, with the output showing the percentage of the instances where Winnow-2 made a guess at all, as well as the percentage of those guesses that were correct. Multiplied, the give the algorithm's total accuracy. Winnow-2's total accuracy of 47.4% was significantly lower than that of Naïve Bayes at 83.8%.

Discussion

Both algorithms got middling scores on the Iris dataset. Neither algorithm was able to accurately learn the Glass dataset. Interestingly the model produced by Winnow-2 does not even include class 4 from the set, and all vectors except one are identical. However, the Glass set resulted in the lowest score for Naïve Bayes, and the second highest for Winnow-2. The House dataset reveals what is likely a bug in the Winnow-2 implementation, as the model for both classes is identical. This error was not caught earlier as the Naïve Bayes algorithm also saw consistent success when classifying this data set. Winnow-2 had the lowest accuracy with the Cancer set while the Naïve Bayes algorithm was able to classify it extremely well. The Soybean set was also very easy for the Naïve Bayes algorithm to classify. Interestingly, the Winnow-2 algorithm attempted to classify every instance in the set, and though its success rate was not proportionally high, it still resulted in it's second highest success rate.

Summary

The data as shown does not display a relationship between the number of classes and the success of either algorithm. Overall, the Naïve Bayes algorithm performed very well, only showing persistent difficulty with one of five data sets, while the Winnow-2 algorithm got both its highest and lowest scores from the datasets that have only two classes. Unfortunately, it is difficult to compare its performance with Winnow-2, due to the high likelihood of a mistake in its implementation and conversion to a multiple-class learning algorithm. Additionally, it is likely that crucial information was lost when converting the continuous and multinomial datasets to Booleans. The range inherent in continuous data may have helped a multiple-class implementation of Winnow-2 learn classes more accurately and may have introduced simplifications that cost the Naïve Bayes algorithm some accuracy.

References

1) Creator: R.A. Fisher Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov) Date: July, 1988
2) Creator: B. German -- Central Research Establishment Home Office Forensic Science Service Aldermaston, Reading, Berkshire RG7 4PN Date: September, 1987
3) Title: 1984 United States Congressional Voting Records Database Source: Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985. Donor: Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu) Date: 27 April 1987
4) O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
5) Michalski,R.S. Learning by being told and learning from examples: an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean desease diagnoiss", International Journal of Policy Analysis and Information Systems, 1980, 4(2), 125-161. Donor: Doug Fisher (dfisher%vuse@uunet.uucp) Date: 1987