Demonstration of an implementation and the performance of two machine learning classifiers: WINNOW-2 and NAÏVE BAYES

David Feldman
9/10/2017

ABSTRACT

The purpose of this experiment was to implement and demonstrate the functionality of two very basic machine learning algorithms: WINNOW-2 and NAÏVE BAYES. After implementing both algorithms in R, we tested their performance on five open source data sets from the University of California Irvine Machine Learning repository (data sets were simplified to include only binary variables as to be compatible with our test algorithms). In the end, WINNOW-2 and NAÏVE BAYES performed identically on 2 of 5 data sets, WINNOW-2 outperformed NAÏVE BAYES on 1 data set, and NAÏVE BAYES outperformed on 2 data sets. Average predictive accuracy for WINNOW-2 was 88% whereas accuracy was 94% for NAÏVE BAYES. This performance fell in line with expectations.

1. Problem Statement and Hypothesis

At a high level, the purpose of this experiment was to make an evaluative comparison between the WINNOW-2 and NAÏVE BAYES algorithms. Because WINNOW-2 can only operate on Boolean classifiers we converted all features of all 5 of our data sets to Boolean form.

Prior to running this experiment, we hypothesized that NAÏVE BAYES would substantially outperform WINNOW-2: NAÏVE BAYES is a powerful (though simple) algorithm that is often used in practice, whereas was WINNOW-2 was developed as a theoretical demonstration, and is not used in practice frequently.

2a. Description: WINNOW-2

The WINNOW-2 algorithm is a technique for building a linear classifier for a target class, from a set of labeled examples. In winnow-2, we iterate through training data with the intent of building the best possible linear classifier matrix – which represents the relative importance of each feature in predicting the classification of the class variable. At a high level, the predictive matrix is developed by setting initial values for each variable in the matrix, and iterating through the training data using the matrix to make predictions. When a correct prediction is made corresponding values in the matrix are 'promoted' and when an incorrect prediction is made corresponding values are demoted. For detailed mathematical representation of WINNOW-2 please see the appropriate reference text. Tuning parameters used for the implementation described herein are:

$$\theta = \frac{n}{2}, \alpha = 2$$

2b. Description NAÏVE BAYES

The Naïve Bayes algorithm is a statistical machine learning algorithm that is based on Bayesian decision theory. The high level of functioning of the algorithm is as follows. For each class value, we calculate the probability (using the training set as our constructive prior) for each value of each independent variable. We then attempt to classify a test sample via taking the product of each independent variable's corresponding probability for each represented class, and the overall class probability. The class corresponding the highest product becomes the predicted class. Please see reference material for a complete mathematical description of NAÏVE BAYES. Tuning parameters used in our implementation are:

$$m = 1\ (laplace\ smoothing), p = uniform\ (.5)$$

3. Experimental Approach

As to modularize this experiment as much as possible, we created 3 separate R scripts for our implementation. *ETL.R* deals with the loading, cleansing, and formatting of data sets. *Implementation.R* implements train and test function for both WINNOW-2 and NAÏVE BAYES. *Run.R* splits data into training and test sets (2/3 , 1/3 respectively) runs our algorithms, and prints trained-state matrices used for prediction, summary statistics and classification results.

At a high level, key items to call out about our approach are:

- For all non-boolean variables we make a rough 'one hot coding' – if a value is less than the variable mean we choose 0, if greater we choose 1. We recognize that this approach is not optimal, should the goal of our experiment be to make our algorithms as predictive as possible.
- Rows with missing values in the breast cancer data set are removed. Missing values in the vote data set are set randomly.

Data sets used for our experimentation are from the UCI Machine Learning Repository and were sourced as follows:

1. Breast Cancer—https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
   This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Class variable divided Malignant = 1, not Malignant = 0.
2. Glass — https://archive.ics.uci.edu/ml/datasets/Glass+Identification The study of classification of types of glass was motivated by criminological investigation. Class variable divided Window glass = 1, not window glass = 0.
3. Iris — https://archive.ics.uci.edu/ml/datasets/Iris
   The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Class variable divided Setosa = 1, not Setosa = 0.
4. Soybean (small) — https://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29
   A small subset of the original soybean database. Class variable = D4 then 1, else 0.
5. Vote — https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records
   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. Class variable Democrat = 1, else 0.

4. Experimental Results and Discussion

All in all, both WINNOW-2 and NAÏVE BAYES performed satisfactorily as classifiers. Naïve Bayes was the better classifier of the two: average predictive accuracy for WINNOW-2 was 88% whereas accuracy was 94% for NAÏVE BAYES. See ***Table 1*** below for specific accuracies on each data set for both algorithms (noting that full summary statistics are available in the program trace).

Data Set	breast cancer	glass	iris	soybean	vote	MEAN
Winnow-2 Accuracy	0.9649	0.9444	0.66	1	0.8483	0.88352
Naïve Bayes Accuracy	0.9649	0.8889	0.96	1	0.8897	0.9407

Table 1: *Predictive accuracy for both algorithms on each data set*

Interestingly, the Winnow-2 algorithm outperformed Naïve Bayes on the glass data set. As it goes, we were attempting to predict whether the glass was window glass or not – which was highly indicated by Mg content, but other variables were much less valuable. Winnow-2 did a good job of assigning other variables lower weight.

Both algorithms did a very good job on Breast cancer and Soybean – indicating that predicting the correct class was relatively simplistic.

The Iris and Vote data sets (especially Iris) were probably more nuanced and complex – hence the better performance by Naïve Bayes. For Iris, Winnow-2 seems have to greatly overweighed Sepal length – and could not recover from this.

6. Summary

All in all we were able to successfully demonstrate the functionality of two very basic machine learning algorithms: WINNOW-2 and NAÏVE BAYES. After implementing both algorithms in R, we tested their performance on five open source data sets from the University of California Irvine Machine Learning repository (data sets were simplified to include only binary variables as to be compatible with our test algorithms).

In the end, WINNOW-2 and NAÏVE BAYES performed identically on 2 of 5 data sets, WINNOW-2 outperformed NAÏVE BAYES on 1 data set, and NAÏVE BAYES outperformed on 2 data sets. Average predictive accuracy for WINNOW-2 was 88% whereas accuracy was 94% for NAÏVE BAYES. This performance fell in line with expectations and proved both algorithms to be effective classifiers.

References

Littlestone, Nick. "Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm." *Machine Learning* 2, no. 4 (April 1988): 285–318. doi:10.1007/BF00116827.

Murphey, Kevin. "Naive Bayes Classifiers." University of British Columbia, October 24, 2006. https://datajobsboard.com/wp-content/uploads/2017/01/Naive-Bayes-Kevin-Murphy.pdf.

"Center for Machine Learning and Intelligent Systems." *UCI Machine Learning Repository*, University of California, Irvine, archive.ics.uci.edu/ml/index.php.