Naïve Bayes and Winnow-2

Amrith Deepak

Abstract: We did an analysis on 3 data sets on House votes, Breast Cancer tests malignant or benign, and Iris data species classification using 2 algorithms, Winnow-2 and Naïve Bayes. Winnow-2 performed slightly better than Naïve Bayes on the House and Breast Cancer data, but the difference wasn't significant. Both performed very well on those 2 datasets and poorly on the Iris Data.

Problem Statement:

We are trying to see how well Winnow-2 and Naïve Bayes algorithms perform on different data-sets to classify data. We tried using 16 different attributes like to predict the class of Congressmen (Democrat/Republican), length and width statistics to predict species of iris plants, and seeing whether breast cancer tests showed malignant or benign based on patient data.

Hypothesis:

I would expect Winnow-2 to perform well on the Congressmen and the breast cancer data, since there are 2 classes. With 3 classes(iris), I feel like the Winnow-2 will perform less well. I feel like Naïve Bayes won't perform as well since the conditional independence assumption may not hold.

Algorithms:

The Winnow-2 algorithm trains data by assigning weights to each attribute. We multiple all the attributes with value 1 by its weights and if the total is less than a predefined theta value, we promote the weights and if it's greater than the predefined theta value, we demote the weights. At the end, we used our trained weights on the testing data and if it is greater than theta, we classify it as class 1 else class 0.

The Naïve Bayes algorithm calculates all the conditional probabilities of each variable given each value of the variable we are trying to predict. It's an application of Bayes theorem. To test the algorithm, we multiply all the conditional probabilities for all values of the variable we are trying to predict and the greatest conditional probability is used.

Experimental Approach:

For each of the three datasets, I split the data into training and testing data. I followed the algorithms described above. For missing values, I randomly filled it in with one of the values. For the iris data-set, since there were more than 2 classes, I did a separate classification for Iris-setosa vs. not iris-setosa, iris-versicolour vs. not iris-versicolour, or iris-virginica vs. not iris-virginica when using the Winnow-2 algorithm. For the winnow-2 algorithm, the training data is used for the getting the weights and for the Naïve Bayes algorithm, the training data is used for getting the conditional probabilities. Those are applied in the testing data.

Results:

The Winnow-2 Algorithm performed very well on the house data where it classified 114 correctly and 12 incorrectly. Similarly, in the cancer data, it classified 101 correctly and 0 incorrectly. However, in the iris data, it performed poorly, where it classified 9 correctly and 20 incorrectly.

The Naïve Bayes Algorithm performed very well on the house data where it classified 112 correctly and 14 incorrectly. Similarly, in the cancer data, it classified 98 correctly and 3 incorrectly. However, in the iris data, it performed poorly, where it classified 9 correctly and 20 incorrectly.

Discussion:

They both performed very similarly, but the winnow-2 performed slightly better in the house and cancer data. Both the models performed extremely poorly on the iris data. It seems like the training data wasn't very well representative of the testing data. Also, the difference in lengths between the 3 species of iris plants weren't as significant as the other data sets, hence the performance was worse. Lastly, there were 3 classes of data, where as in the other 2 data sets, there were only 2 classes. Naïve Bayes assumes conditional independence among attributes and sepal length, sepal width, petal length, and petal width don't satisfy those criteria.

Summary:

Both the algorithms Winnow-2 and Naïve Bayes Algorithm generally performed very well on the House Data and the Breast Cancer data but performed very poorly on the Iris data. These are both classification algorithms and were trained with approximately 2/3 of the data and tested with approximately 1/3 of the data.

References:

https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf
http://www.cs.cmu.edu/~ninamf/ML10/winnow.ps