Naïve Bayes and Logistic Regression

Amrith Deepak

Abstract:

The breast cancer data showing tumors were malignant, glass data showing type of glass, Iris data showing species of Iris plants, soybean data examining the class of soybeans, and vote data examining the party of house congress votes were analyzed using Naïve Bayes and Logistic Regression. All the datasets performed better with logistic regression, although the difference was most pronounced in the soybean dataset.

Problem Statement:

This project is examining how well logistic regression works and comparing it to Naïve Bayes. These algorithms are being evaluated on how they perform on different datasets to classify data. 9 attributes are being used to predict if a tumors are malignant, 9 attributes are being used to predict type of glass, length and width statistics to predict species of iris plants, 35 attributes are used to predict type of soybean, and 16 attributes are used to predict whether the Congress vote is Republican or Democrat.

Hypothesis:

It is predicted that Logistic Regression will perform better on all the datasets since in most of these datasets the conditional independence assumption does not hold, which would cause Naïve Bayes to not perform well. In the house data and the iris data, conditional independence looks to be violated more, so Naïve Bayes would be expected to perform worse there. The assumptions of logistic regression like observations being independent of each other, little or no multicollinearity among independent variables hold here. The assumption of independent variables linearly related to the log odds sometimes may not hold, but overall logistic regression is still expected to outperform Naïve Bayes.

Algorithms:

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

Logistic regression uses a logarithmic function to model the probability of a certain class or event occurring. There is a sigmoid function which is used to compute the class and gradient function which calculates the gradient descent. For the logistic regression function, since there are more than 2 classes, for every class weights are calculated and the class with the highest

weight is chosen. The logistic regression function iterates until the weights do not change. The weights change based on the gradient function and learning rate specified as an input to the function.

Experimental Approach:

For each of the five datasets, the data was split into training and testing data using 5-fold cross validation. The algorithms described above were followed.

For each of the three datasets, the variables were changed to binary numerical variables based on whether they were greater than or less than the mean. Missing values from the house votes dataset were randomly filled in. For the breast cancer data, missing values were removed since there were small in quantity.

Results:

Dataset	**Naïve Bayes Performance**	**Logistic Regression Performance**
Breast Cancer	0.6485	0.9029
Glass	0.2169	0.9029
Iris	0.3333	0.74
Soybean	0.2241	1.0
Vote	0.5698	0.9605

The Naïve Bayes performed moderately well on the breast cancer data and house votes data where it classified with accuracies 0.6485 and 0.5698 respectively. In the glass, iris, and soybean data, it performed with 0.2169, 0.3333, and 0.2241 accuracy respectively.

The logistic regression performed very well on the breast cancer, glass, soybean, and vote data where the performance was greater than 90%. It performed moderately well on the soybean data with 0.74 accuracy. On the soybean data, it predicted with 100% accuracy.

In all cases, the logistic regression had better performance than Naïve Bayes. The largest difference was in the soybean data where Naïve Bayes performed with 0.2241 accuracy, where as logistic regression has 1.0 accuracy. On the breast cancer and vote data, Naïve Bayes performed better so the difference was less, and on the iris data, logistic regression performed worse, so the difference was less. However, in all cases there was a significant difference in performance.

Discussion:

All the datasets performed significantly better for logistic regression than Naïve Bayes. In most cases, the Naïve Bayes performance was under 50% but in logistic regression, the performance was over 90%. One big reason for the weaker performance of Naïve Bayes is that it assumes

conditional independence which may not hold for many of the datasets. For the breast cancer and vote data, Naïve Bayes performed better than the glass, iris, and soybean data. The breast cancer and vote datasets only had 2 classes, where as the other datasets had more than 2 classes.

The logistic regression model correctly classified all instances in the soybean dataset, but that was the smallest dataset. In the house vote data, it had over 96% accuracy. In the breast cancer and glass dataset, it performed with over 90% accuracy and 74% accuracy in the iris data. The iris data had variables like sepal length, width, and petal length, width which are correlated with each other. For these data sets, the logistic function was a good fit, hence the logistic regression model produced accurate results.

Summary:

All the datasets performed better with logistic regression, although the difference was most pronounced in the soybean dataset. Naïve Bayes performed better in the breast cancer and vote dataset when there were only 2 classes, but still considerably worse than Naïve Bayes. The code trained using 5-fold cross validation where each time 20% of the data was used for testing and 80% for training.

References:

Works Cited

Kaur, Gurneet, and Neelam Oberai. "A REVIEW ARTICLE ON NAIVE BAYES CLASSIFIER WITH VARIOUS SMOOTHING TECHNIQUES." *A REVIEW ARTICLE ON NAIVE BAYES CLASSIFIER WITH VARIOUS SMOOTHING TECHNIQUES*, Department of Computer Science, MAHARISHI MARKANDESHWAR UNIVERSITY, 10 Oct. 2014, pdfs.semanticscholar.org/1c41/b7b724e1245201c895160fb46cdd84dca809.pdf.

Peng, CHAO-YING JOANNE, et al. "An Introduction to Logistic Regression Analysis and Reporting." *An Introduction to Logistic Regression Analysis and Reporting*, Indiana University-Bloomington, 2002, datajobs.com/data-science-repo/Logistic-Regression-[Peng-et-al].pdf.

Datasets:

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

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

https://archive.ics.uci.edu/ml/datasets/Iris

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

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