COMP 551 Applied Machine Learning Winter 2020
Logistic Regression and Naive Bayes

Smith, Brandon
McGill University
brandon.smith@mail.mcgill.ca

Dawood, Sully
McGill University
sully.dawood@mail.mcgill.ca

Yang, Danning
McGill University
ydn2012@hotmail.com

I. Abstract

As specified by the instructions, this project focuses on implementing Logistic regression and Gaussian Naive Bayes as probabilistic machine learning classifying algorithms. As specified, we worked on the Ionosphere and Adult data sets, with Haberman's Survival Data and Wisconsin's Breast Cancer database as additional data sets. The results of each model were cross-referenced with a script written for k-fold cross validation. Although the parameter of the k-fold could be changed, a 5-fold cross validation was run on each of the four data sets to verify the accuracy of our models. In general, we found that the Gaussian naive Bayes model was faster at training, predicting, and validating data than our logistic regression model. Furthermore, we also found that the Gaussian Naive Bayes model had an average accuracy rate of 0.8144 across the four data sets, whilst the Logistic Regression model had an average accuracy rate of 0.7586. However, there were certain considerations that we believe had played a part in the slightly lower accuracy of the logistic regression model, such as limiting the number of iterations for large data sets. These shall be explained further in the Results section. However, overall the results show that although the Gaussian Naive Bayes had an accuracy rate of over 80 percent, the logistic regression model should be able to catch up to it for large data sets with large number of instances (n). Additionally, the choice of learning rate was shown to be dependent on the data-set the model is being trained on. Where i is such that the learning rate = 0.1^{7-i}, the optimal i was 5, 7, 1, 4 for the Ionosphere, Adult, Haberman, and Breast Cancer data sets respectively.

II. Introduction

The two machine learning models examined in this paper are Gaussian Naive Bayes and Logistic Regression. The basic concept behind Logistic Regression is to encapsulate a desired function called a "logit" in another function called a "logistic function" which restricts the logit such that it maps to values between 0 and 1. If the logistic function predicts a value of $<$ 0.5 it would be classified as 0, otherwise if it predicts a value of $>=$ 0.5 it would be classified as 1. From there the logistic function can be utilized in defining what's known as a cross entropy loss function which can be simplified to a cost function. From there the Gradient Descent method can be used to find the weights for predicting if a feature vector would be classified as 0 or 1 by evaluating the logistic function with said weights. Gradient Descent works by taking the derivative (partial derivatives evaluated at the current weights) of your cost function, multiplying it by the "learning rate" constant, and subtracting the result from the current set of weights. Repeat the process for finite steps until "learning" appears to have come to a halt.

Finally, Naive Bayes works on the principle of Bayesian Probability. Unlike Logistic Regression, Naive Bayes does not work on the concept of iterative learning through methods like Gradient Descent, but instead relies solely on probabilistic maths. Using the training portion of the data-set you can "learn" the prior and the likelihoods of each feature-vector, and then these learned values can be used to calculate the

probability of a given feature vector being classified one way or another ($< 0.5 = 0$, and $>= 0.5 = 1$).

Due to this purely probabilistic math approach Naive Bayes should be expected to run faster than Logistic Regression, but additionally Naive Bayes should produce more accurate results on smaller data-sets with Logistic Regression eventually catching up to or surpassing it. Logistic Regression should also run at different speeds for different learning rates, with the optimal learning rate value being unique to each data-set.

One of the papers that is certainly worth mentioning is the paper on the comparison of discriminative vs generative classifiers with respect to logistic regression and naive Bayes by Andrew Y. Ng and Michael I. Jordan.[8] Not only does the paper analyze 15 UCI machine learning data sets with Naive Bayes and Logistic Regression, it also includes a thorough discussion of two of the data sets used in our study, the Ionosphere and Adult data sets.

Furthermore, the paper serves as a robust reference point as it addresses the tendencies of asymptotic error to be lower (on average) for logistical regression, while convergence of said error might be faster but higher for naive Bayesian methods. It also solidifies that certain theoretical predictions hold, especially as the number of instances (m) is increased. In other words, it confirms that discriminative logistic regression performs better than naive Bayesian methods when instances are sufficiently large for training purposes, but when it comes to initial classification, naive Bayes may still perform better, especially with smaller data sets.[8]

In addition to the aforementioned paper, another one titled "Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence" by Witteveen et. al provided insight given that the third and forth data sets used are for cancer survival rate and tumor size prediction respectively. The paper statistically compares logistic regression and Bayesian networks, and states how generally logistic regression outperformed Bayesian networks when it came to predicting locoregional recurrence (LRR) and second primary (SP) breast cancer risk. The primary statistic to gauge which method outperformed was the c-statistic. Nevertheless, although there was logistical regression out performance for the breast cancer data (2003-2006, Netherlands Cancer Registry, N = 37,320), the paper stated that predictive differences between the Bayesian and logisitic methods were still relatively small.[9]

III. Datasets

In our project we worked on four data sets: Ionosphere, Adult, Haberman, and Wisconsin Breast Cancer database. We cleaned up the data sets by removing any duplicate instances and/or any instances with a "?" for any of the instances' respective features. Some clean ups were unique to data sets, for example we removed all instances of data for the Ionosphere data set where readings were constantly negative for all features of an instance and also started with a zero. In addition to this, our DataAnalze.py file breaks down each data set into four different classes, each class named after each of them respectively, and each class has methods that respectively clean the data and analyze it. Furthermore, we return the cleaned features and class data in a matrix/array form to help manipulation in our MLModel.py.

The first data set, the "Johns Hopkins University Ionosphere database", has 34 radar data continuous attributes/predictor variable, with the predictive task being to classify "good" and "bad" signals. The second data set, "Adult.data", is more extensive in terms of instances. It has 14 attributes/predictor variables and the predictive task is to determine a person's income and whether it is under or over $50,000.

In addition to the above two data sets, we also used "Haberman's Survival Data", which recorded certain patients' numerical attributes and their cancer survival status as the class attribute. Lastly, the "Wisconsin Breast Cancer

Figure 1: *Adult - Co-variance Matrix Heat Map*

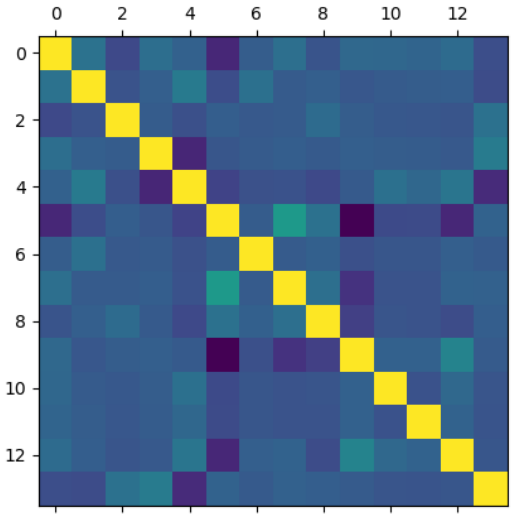
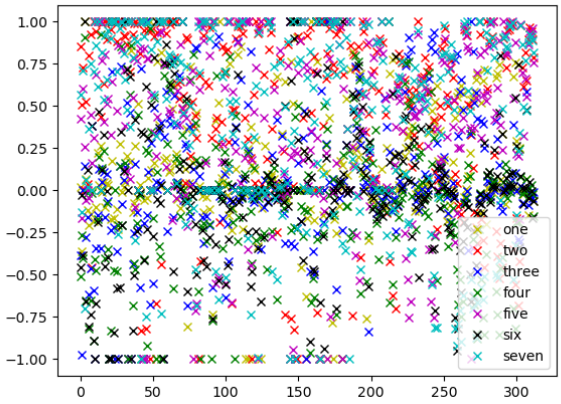


Figure 2: *Ionosphere - Scatter Plot*



database" was the final data set we used, which contained 10 numerical attributes, with the classification task being to determine whether a tumour was "benign" or "malignant".

The distribution of features from the Haberman data set follows a Gaussian distribution, with a hint of skewness. It is important to note that the data was normalized using the density parameters.

Furthermore, we also plotted heat maps, scatter plots, and Co-variance matrices to better understand the data.

The figures above are just some of the statistics run to better understand the data. The ones presented are ones that were found relevant. For example, in figure 2 we find that there seems to be strong correlation between variables 5 and 7 (0.3677), which are respectively education and occupation; this makes sense from a logical standpoint. Additionally, figure 3 shows a scatter plot analysis of seven signals we arbitrarily chose to better understand the spread of the Ionosphere data. There seems to be a greater likelihood of positive signals. Lastly, in figure 4 we also found certain variables to heavily co-vary in the Breast Cancer data set. For example, variable 2 and 3 (Clump thickness and Uniformity of cell size) showed

Figure 3: *Breast Cancer Wisconsin - Co-variance Matrix*

	0	1	2	3	4	5	6
0	1.000000	-0.056350	-0.041396	-0.042221	-0.069630	-0.048644	-0.099248
1	-0.056350	1.000000	0.642481	0.653470	0.487829	0.523596	0.593091
2	-0.041396	0.642481	1.000000	0.907228	0.706977	0.753544	0.691709
3	-0.042221	0.653470	0.907228	1.000000	0.685948	0.722462	0.713878
4	-0.069630	0.487829	0.706977	0.685948	1.000000	0.594548	0.670648
5	-0.048644	0.523596	0.753544	0.722462	0.594548	1.000000	0.585716
6	-0.099248	0.593091	0.691709	0.713878	0.670648	0.585716	1.000000
7	-0.061966	0.553742	0.755559	0.735344	0.668567	0.618128	0.680615
8	-0.050699	0.534066	0.719346	0.717963	0.603121	0.628926	0.584280
9	-0.037972	0.350957	0.460755	0.441258	0.418898	0.480583	0.339210

Figure 4: *Breast-Cancer Training*

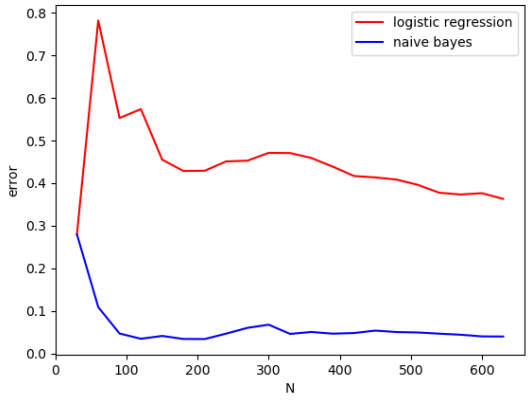
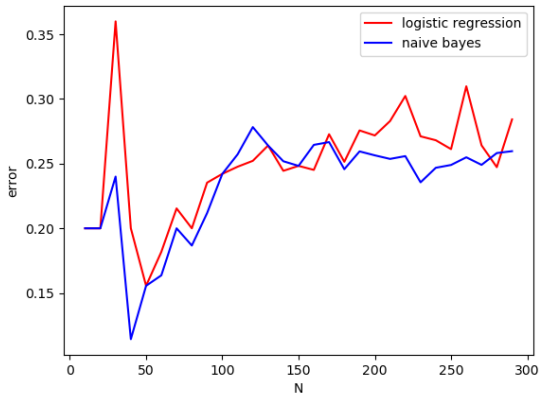
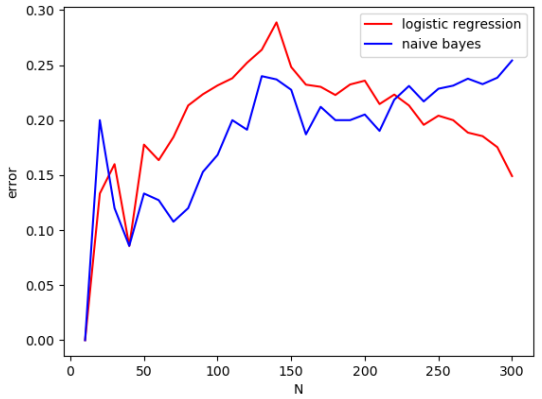


Figure 5: *Haberman Training*



a co-variance of 0.907, which is perhaps an understandable abnormality.

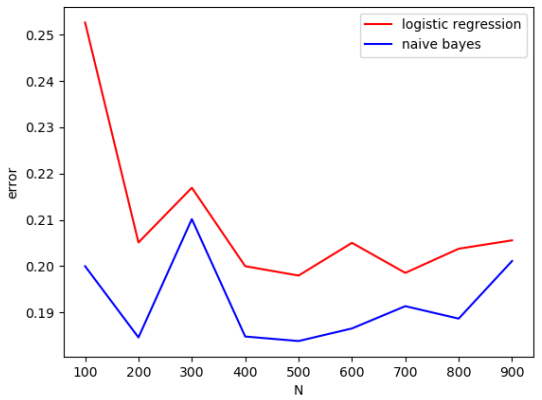
Figure 6: *Ionosphere Training*



IV. Results

Over the four data sets, our Gaussian naive Bayes method had an average accuracy of 81.44 percent, whilst the logistic regression had an accuracy of 75.86 percent. Both of these accuracies were gauged by our 5-fold cross-validation. The results of our logistic regression model for the data sets were as follows: Ionosphere: 0.849, Adult: 0.785, Haberman: 0.75, Wisconsin Breast Cancer: 0.65. The results of our Gaussian naive Bayes model were as follows: Ionosphere: 0.737, Adult: 0.809, Haberman: 0.75, Breast Cancer Wisconsin: 0.96. However, it is important to note that due to the large number of instances in the Adult and Haberman data sets, we added a maximum iteration variable to the Gradient Descent function in our Machine Learning model (MLModel.py). This was a decision made to trade accuracy for efficiency, as it helped garner results across the data sets at a faster pace. This variable was set to 1000 for the two Adult and Haberman data sets, but had it been increased the logistic regression model would have garnered results closer to the naive Bayes model.

Figure 7: *Adult Training*



Another aspect that was tested on the four data sets was the variance of logistical regression performance with respect to learning rates

Figure 8: *Running Time of Adult for Varying Learning Rates*

Figure 9: *Running Time of Breast-Cancer for Varying Learning Rates*

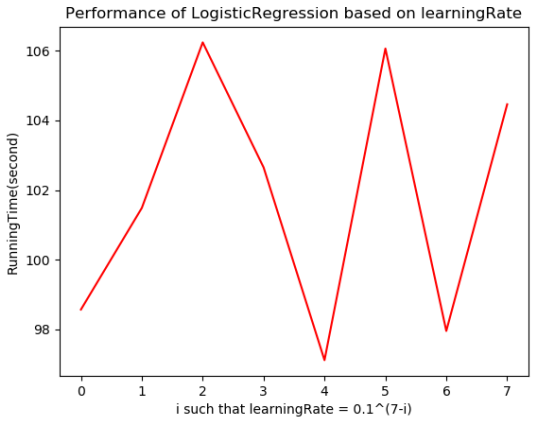


Figure 10: *Running Time of Haberman for Varying Learning Rates*

Figure 11: *Running Time of Ionosphere for Varying Learning Rates*

and their respective convergence. The findings were that the optimal learning rates for each data set varied. Where i is such that the learning rate = 0.1^{7-i}, the optimal i was 5, 7, 1, 4 for the Ionosphere, Adult, Haberman, and Breast Cancer data sets respectively.

There was also a comparison made between the error rate (i.e. validation accuracy) of the two algorithms and the number of instances we trained them on. For the Ionosphere data set, we see that the error rate is initially lower for naive Bayes, but as the number of instances increase, the logistic regression algorithm catches up and even out performs it. For the Haberman data set, we see that both algorithms have very similar error rates as n is increased, but one notable aspect is the out performance of naive Bayes for lower number of instances. Similarly, we see that for the Breast Cancer data set, not only does naive Bayes out perform initially, but it consistently maintains the lower error rate. Lastly, we also find that the naive Bayesian theory holds for the adult data set, as it consistently out performs as well (although logistical regression catches up as n increases).

V. Discussion and Conclusion

In conclusion, the theoretical assumptions about Naive Bayes out performing Logistic Regression seems to strongly hold for the Adult,

Haberman and Wisconsin Breast Cancer data sets. For the Ionosphere data set it seems to still hold, though not as strongly; however, Naive Bayes still out performs until approximately n = 220. Another interesting caveat is the effect of learning rates in our logistic regression. We found that through trial and error, a unique efficient learning rate could be found that would optimize running time for each data-set. Lastly, we also found that running a 5-fold cross validation helped verify our predictive results. A caveat for future research could be to look at which subset of features are statistically significant in predicting regressive values, and then basing them as our primary input parameters/weights (and vice versa for non-statistically significant ones). Furthermore, a hybrid between the two classifiers would be ideal, where one could combine the best of both worlds - efficiency and accuracy.

VI. Statement of Contributions

All members actively participated in the coding and report write up process. Work was steadily built up over the course of a few weeks, with all team members meeting regularly to ensure consistency. Team also went to multiple TA office hours together to aid understanding of course material.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC.

[2] Kevin P. Murphy. *Machine Learning : A Probabilistic Perspective*. MIT Press.

[3] Ron Kohavi. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, to appear.

[4] Dr. William H. Wolberg (physician). *University of Wisconsin Hospital, original source for breast-cancer-wisconsin dataset*.

[5] O.L. Mangasarian, R. Setiono, and W. H. Wolberg. *Pattern Recognition Via Linear Programming: Theory And Application To Medical Diagnosis*. 1990.

[6] Tjen-Sien Lim (limt@stat.wisc.edu). *Haberman's Survival Data*.

[7] Vince Sigillito (vgs@aplcen.apl.jhu.edu). *Johns Hopkins University Ionosphere database*.

[8] Andrew Y. Ng and Michael I. Jordan. *On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes*. https://pdfs.semanticscholar.org/9092/9a6aa901ba958eb4960aeeb594c752e08369.pdf?fbclid=IwAR3h-9t-PIRXNM0zQJsQWwgEUPalZU7R-GvDWCK5Fizq3iit7IruNBuyKpU.

[9] Annemieke Witteveen et al. *Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence*. https://journals.sagepub.com/doi/pdf/10.1177/0272989X18790963?fbclid=IwAR20nM62_mvZI8MMLrq_EcXV1OR3t7_uEln6K4A3oEylhl37SrpCwIpn2aE&.