Comparing Classification Techniques: Logistic Regression and Naive Bayes

Eric Anderson, Ali Tapan, Yawan Xu
McGill University
COMP 551 - Applied Machine Learning
Miniproject 1
Winter 2020

Abstract— **This work applies two classification models, logistic regression and naive Bayes, to four distinctive datasets.**

1) **The *Ionosphere Data Set*[1] where the goal is to predict whether a radar return from ionosphere is good, i.e., shows evidence of structure in the ionosphere, or bad.**
2) **The *Adult Dataset* or *Census Income*[2] dataset, where the goal is to predict whether income exceeds $50k a year based on census data.**
3) **The *Breast Cancer Wisconsin Data Set*[3] which comprises of clinical cases of medical diagnosis, where the goal is to predict whether the tumor is benign or malignant.**
4) **The *Wine Quality Data Set*[4] which comprises of the physiochemical attributes of red wine samples from Portugal. The goal is to predict the quality of the wine on a scale of 1-10.**

We analyzed and compared the performance of the two models, and found that the logistic regression approach achieved on average a 2.7% higher accuracy than naive Bayes. Moreover, their performance is similar based on the confusion matrix that they generated. Further, we found that logistic regression is significantly slower to train than naive Bayes, especially on large datasets such as the *Adult Data Set*.

I. Introduction

A. Background

Logistic Regression (LR) and Naive Bayes (NB) are two commonly used classification models in machine learning. The first is a variation of linear regression, where a model calculates the dependent variable based on the independent variable. The latter is based on joint distribution, and assumes that the features are independent, hence the "naive" in its name.

B. Task Summary

In this work, we implemented LR and NB and compared their performance on four datasets. We first preprocessed and analyzed the data, then implemented the models from scratch using Python. Finally, we ran the experiments and compared the accuracy of both models on the four datasets using 5-fold cross validation. We also tested the performance of LR with different learning rates, as well as evaluating the two models with different training sizes.

C. Important Findings

We found that learning rate is an important hyperparameter that has a direct impact on the accuracy of our model. For all four datasets, we obtained that increasing the learning rate from 0.0005 to 0.1 improves the accuracy. However, this improvement stalls or reverses as the learning rate goes beyond 0.1. Furthermore, we found that smaller learning rates take longer to reach an optimum, while the cost converges much faster with higher learning rates.

II. Datasets

A. Ionosphere

The *Ionosphere Data Set* contains 351 instances with 34 features each. This data was used to predict whether radar output from an ionosphere was 'good' or 'bad'. A bias column was added to the dataset for the LR.

B. Adult

The *Adult Data Set* is a large dataset which comprises of 48842 instances from the Census bureau. The task is to predict whether a given adult makes more than $50,000 a year based on a mix of discrete and continuous attributes. In the LR program, the data was scaled on a 0 to 1 range to improve performance and calculation speed. Out of the 14 features, 8 were categorical. We converted these variables to numerical data using one-hot encoding for LR and label encoding for NB.

C. Breast Cancer

The Breast Cancer Wisconsin Data Set comprises of 699 clinical cases of medical diagnosis for breast cancer. The ID column was dropped due to its irrelevance to the prediction task, which is determining whether the tumor is malignant (1) or benign (0). Furthermore, 16 samples were removed due to missing value in one of the remaining 9 features, leaving us with 683 training data.

D. Wine Quality

The *Wine Quality Data Set* comprises of 1599 instances of wine samples with 11 physiochemical attributes, and a quality score ranging from 0 to 10. This is the only dataset on which we performed multiclass classification. The training data was scaled to a range of 0 to 1 to improve the LR performance.

Figure 1 shows the class distribution of the quality scores across all wine samples. As we can see, the distribution is not balanced, as there are more wines scoring a middle score of 5-6. There is also no score below 3 or above 8.

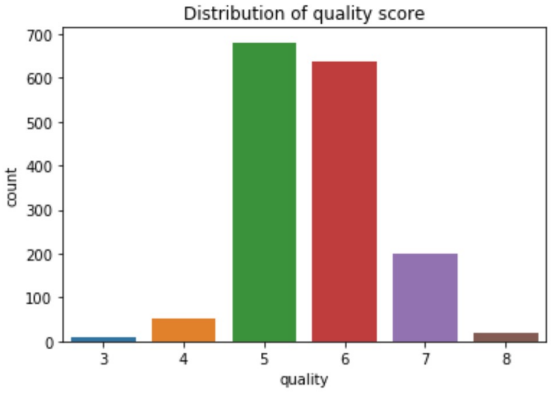

Fig. 1: Class distribution of wine quality.

III. Results

Table 1 shows the results of both models on all four datasets. LR outcompetes NB for all except the breast cancer dataset. Considering the small size of the latter dataset, this was not surprising, as it is generally the case that generative models such as NB work better on smaller datasets.

Dataset	Logistic Regression	Naive Bayes
Ionosphere	87.7%	82.3%
Adult	84.7%	82.1%
Breast Cancer	95.8%	96.3%
Wine Quality	57.7%	54.5%

TABLE I: Mean accuracy of LR and NB based on datasets.

The plots for the accuracy on train/validation set as a function of iterations of gradient descent for each dataset can be found in the appendix (Figure 9-12). From these graphs we can see the effect that the amount of training data has on how consistently model accuracy is able to increase and converge near its maximum.

The wine quality dataset scored a mean accuracy of below 60% on both models. We hypothesize that the poor results are due to the nature of the task, which is multiclass classification, and the fact that the dataset is heavily imbalanced, as seen from figure 1. Considering the simplicity of our models, the poor accuracy is not surprising. In order to do better, we would have to use the one-vs-rest training scheme, or change the loss function in LR to cross-entropy loss[5]. Furthermore, we could work to have a more even distribution by resampling methods. Nonetheless, we can see from figure 5 and 6 (in Appendix) that our vanilla LR and NB were still able to predict near true scores even when it was wrong.

To test the effect of different learning rates for gradient descent, we ran the LR model with learning rates of 0.005, 0.001, 0.005, 0.01, 0.05, and 0.1. We found that accuracy improves as the learning rate increases from 0.0005 to 0.01, then stalls or even reverses on some dataset when the learning rate is larger than 0.01. Figure 2 shows a plot of the mean accuracy against different learning rates for the Adult dataset.

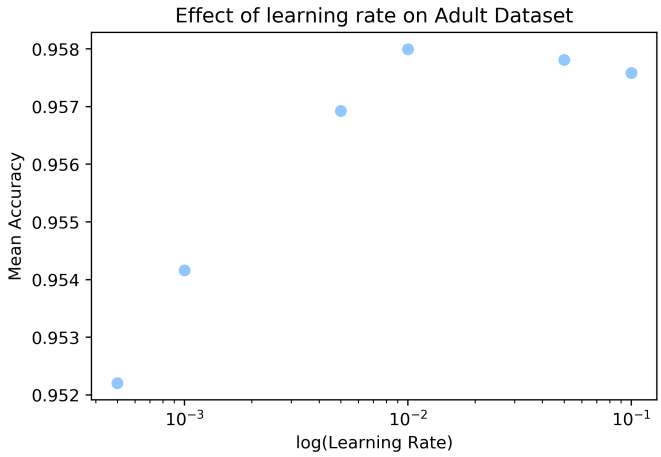

Fig. 2: The effect of different learning rates on prediction accuracy of the Adult Dataset.

The different learning rates also converged at different rates, with lower learning rates taking longer to reach an optimum, if at all, while higher learning rates could overshoot the minimum. Figure 3 shows how the gradient descended with the smallest learning rate, 0.001, while figure 4 shows how it descended with a larger learning rate.

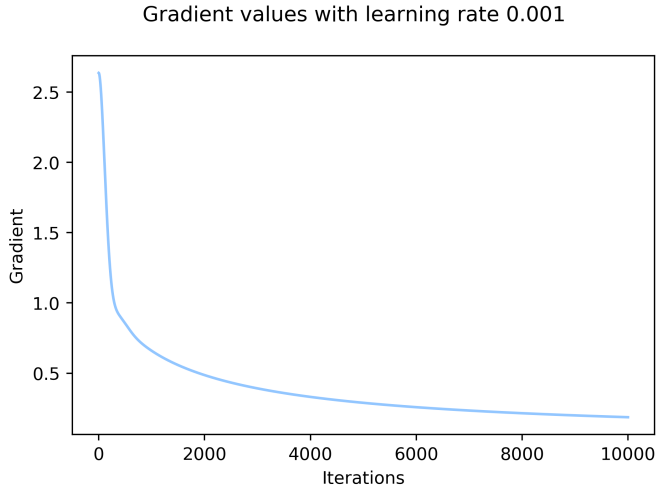

Fig. 3: Gradient descent with a small learning rate.

Additionally, changing the LR training set size from 50% of the total to 90% of the total set led to a minimal increase in accuracy among all sets except for the ionosphere data. This may be due to the smaller size of the ionosphere set, as more available training data would have made training substantially more accurate if there were initially relatively few training samples.

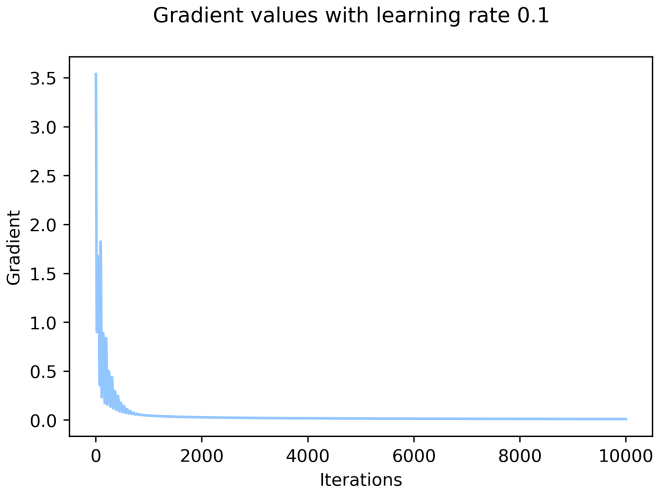

Fig. 4: Gradient descent with a larger learning rate.

The NB algorithm used for the datasets was Gaussian Naive Bayes which used equation (1) and (2) to predict posterior probabilities from the data. The values required to calculate the posterior probabilities were the mean μ and variance σ of the class.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \cdot exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \quad (1)$$

To predict the outcome, the class with the highest probability was chosen.

$$y = argmax_y(\sum_{i=1}^{n} log(P(x_n|y))) + log(P(y)) \quad (2)$$

IV. Discussion and Conclusion

From these results, we can see that relatively simple models are able to achieve accuracy above 80% on data sets of a variety of sizes. For LR, the size and scale of the input data significantly affected how quickly the model performed its calculations, as large numbers and large sets took substantially longer to compute optimal values for. This made data scaling and early stopping valuable in limiting the time spent training the model for negligible accuracy increases. Furthermore, both LR and NB models showed to be not sufficient to predict the outcomes of multiple class data sets. This can be observed from Figure 4 and Figure 7 where both models have different actual and

predicted values. With mean accuracy of 57.7% for LR and 54.5% for NB shows further improvements are required to evaluate such datasets. These poor accuracies could be improved with more balanced data, as both confusion matrices in figures 5 and 6 for LR and NB show similar difficulty predicting classes with fewer samples. (See Appendix)

V. Statement of Contributions

- **Eric Anderson**: Built and implemented the Logistic Regression algorithm, ran the experiments and found the best hyperparameters. Included regularization and momentum, constructed a kfold cross validation for the logistic regression, and scaled data. Also contributed to part of the report.

- **Ali Tapan**: Built and implemented the Naive Bayes algorithm on all four datasets, ran the experiments and contributed to part of the report.

- **Yawan Xu**: Cleaned and did exploratory data analysis on the datasets, contributed to part of the logistic regression modelling and implementation, and wrote part of the report.

References

[1] Sigillito V. (1989). Ionosphere Data Set [https://archive.ics.uci.edu/ml/datasets/ionosphere]. Johns Hopkins University, Johns Hopkins Road Laurel, MD 20723

[2] Kohavi R. and Becker, B. (1996). Adult Data Set [https://archive.ics.uci.edu/ml/datasets/Adult].

[3] Wolberg, W. (1992). Breast Cancer Wisconsin Data Set [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+ Wisconsin+%28Original%29]. University of Wisconsin Hospitals, Madison, Wisconsin, USA

[4] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[5] Kumar, A. (2019, January 10). Logistic Regression for Multi-class Classification with Example in Python. Retrieved from https://acadgild.com/blog/logistic-regression-multiclass-classification

Appendix

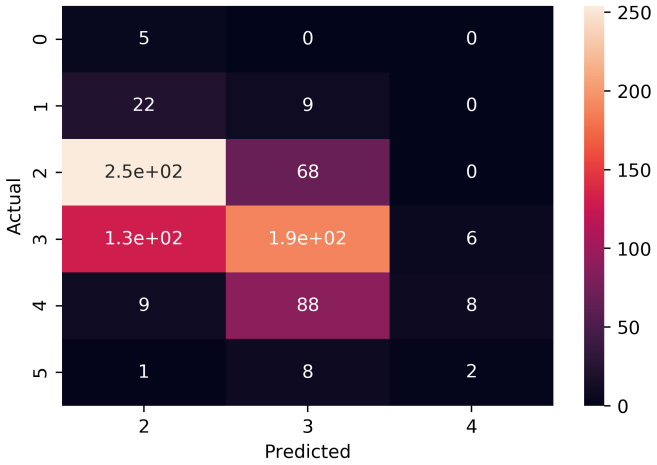

Fig. 5: LR confusion matrix on the wine data set.

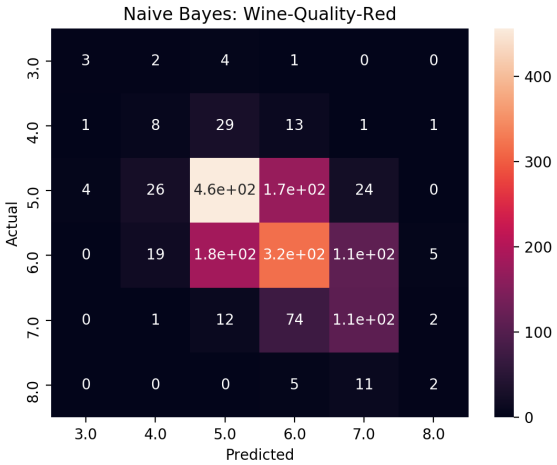

Fig. 6: Naive Bayes confusion matrix on the wine dataset.

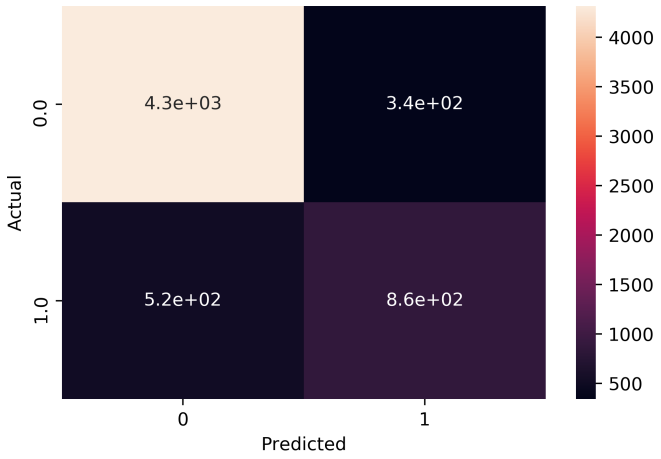

Fig. 7: Logistic Regression confusion matrix on the adult dataset.

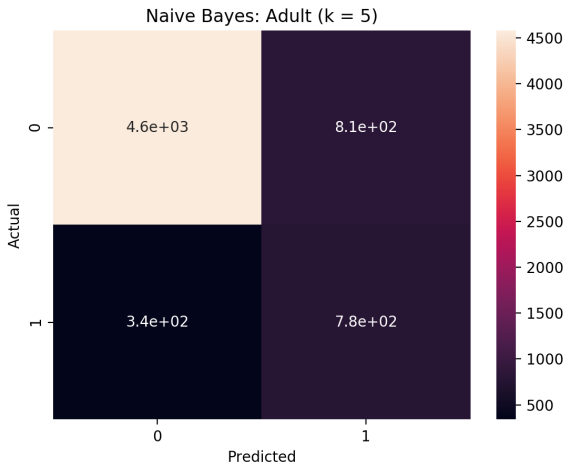

Fig. 8: Naive Bayes confusion matrix on the adult dataset.

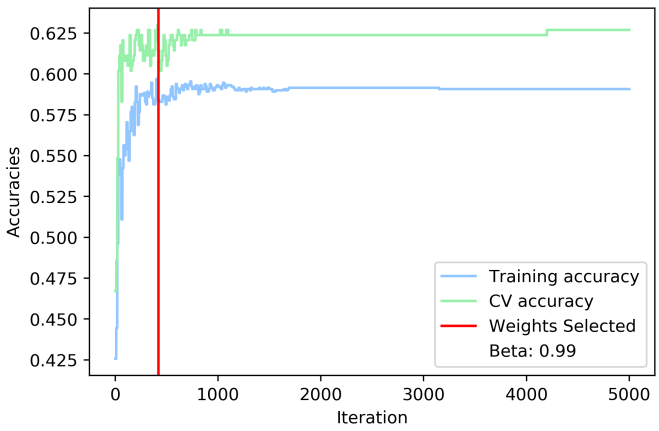

Fig. 9: Accuracy of LR on the Wine dataset as a function of iterations.

Fig. 10: Accuracy of LR on the Adult dataset as a function of iterations.

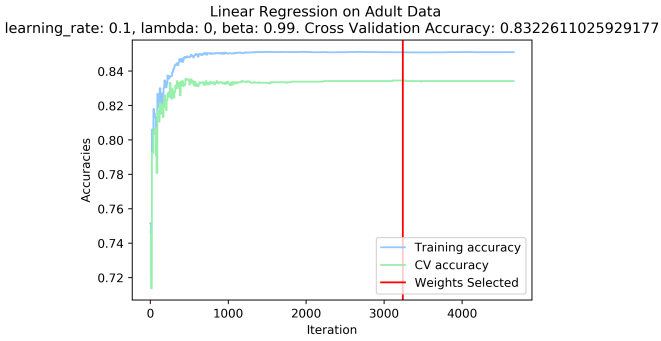

Fig. 11: Accuracy of LR on the Breast Cancer dataset as a function of iterations.

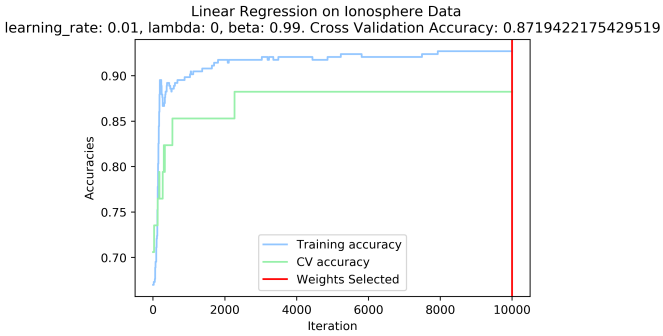

Fig. 12: Accuracy of LR on the Ionosphere dataset as a function of iterations.