Project 1 - Group 62

Mamoun Benchekroun, Ines Moreno, Julien Mounthanyvong

February 12, 2020

1 Abstract

In this mini-project we investigated the performance of two different linear classification models: Logistic Regression and Gaussian Naive Bayes. Logistic regression is a classification algorithm that uses the logistic sigmoid function to predict the probability that some input belongs to a class. On the other hand, Gaussian Naïve Bayes follows a probabilistic approach: it applies Bayes' theorem assuming independence between features and computes the maximum likelihood some input has of belonging to a specific class. We compared the results of these two models on four datasets (Abalone, Adult, Ionosphere and, Iris)and we found that generally, the Naïve Bayes classification not only resulted in a better accuracy than the Logistic Regression one but it also ran faster.

2 Introduction

Naive Bayes and Logistic Regression are two of the most common models used for classification in machine learning. In this project, we compared these two approaches for classification on four different datasets. Naive Bayes, is a generative model that learns the joint probability, $p(x, y)$, of the inputs x and the label y and makes its predictions using Bayes Theorem to calculate the posterior probability, $p(y = c|x)$ and pick the most likely class. We decided to implement Gaussian Naive Bayes that is the most commonly used since it deals with continuous data and uses the equation for a normal distribution parameterized by the mean and the variance. Alternatively, discriminative models such as Logistic Regression, model the posterior probability $p(y|x)$ directly, or learn a direct map from inputs x to the class labels.[4] We tested our models on four distinct datasets from the UCI machine learning repository [2]. First, the Abalone dataset that is used to predict the age of an abalone from physical measurements. Second, the Adult dataset which aims to predict whether a person makes over 50K a year. Third, the ionosphere dataset that predicts whether a radar return from ionosphere is 'good' or 'bad' and finally, the iris dataset which is used to predict the Iris species based on some physical measurements. Contrary to the results obtained by Andrew Ng. and Michael Jordan [4], we observed that the Naive Bayes classifier performed better on most datasets (except for the Iris training data). We believe this is due to the size of our data.

3 Datasets

3.1 Abalone dataset

The goal of this dataset is to predict the age of a abalone (a kind of mollusk) based on measurements of some physical attributes (for example, the number of rings it has when its shell its cut through the cone) as well as other features such as weather patterns or location. The data was collected from the abalone population in Tasmania, Australia in 1994 and it counts with 4177 instances, 8 attributes and 29 classes. During pre-processing, we used one-hot encoding to encode the 'Sex' attribute and we replace all the '?' values by NaN to easily remove them afterwards. In order to have a binary task, we separated the instances in two classes: class 0 for those having a number of 'Rings' below 10 and class 1 for the rest of them. The physical attributes of the abalone are highly correlated but that is to be expected since they are describing the same organism for each instance. We judged it was not worth dropping the features in this case.

3.2 Adult dataset

The goal of this dataset is to predict whether a person's income exceeds \$50K/yr based on census data. The data was extracted from the 1994 Census database by Barry Backer. This dataset counts with 48842 instances and 14 attributes. For the pre-processing, we decided to drop the 'Education' attribute since it was redundant with the 'Education-Num' one and we replace all the '?' values with NaN in order to easily remove them afterwards. Finally, we used one-hot-encoding to encode categorical attributes such as 'Work class', 'Marital Status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Native Country' and 'Salary'. This dataset was also regularized in order to accommodate the 'Capital Gain' and 'Capital Loss' features since they had very high peaks and a lot of zeros which made training hard to implement.

3.3 Ionosphere dataset

The goal of this data set is to predict whether a radar return from ionosphere is 'good' or 'bad'. The radar data was collected by the Space Physics Group of the Johns Hopkins University Applied Physics Laboratory in Goose Bay, Labrador. The system counts with 16 high-frequency antennas with a total power of 6.4 kW and it targeted free electrons in the E- and F- layers of the ionosphere. When the radar return showed some type of structure in the ionosphere it was labeled as 'good' and 'bad' if it didn't. [5] This dataset counts with 351 instances, 34 attributes and 2 classes. When pre-processing, we decided to drop the second column, as it was only zeros and we used one-hot-encoding to encode the 'Result'. We also removed the columns for those features that were highly correlated (>0.98) in order to avoid multicollinearity, which usually yields in linear models to solutions that are extremely varying [3] and possibly numericallly unstable. [1] For this dataset, we presume that the most important thing would be to avoid any false negative as we do not wish to miss a ionosphere with the desired characteristic. Therefore, it might be interesting to look at the recall rather than just the accuracy.

3.4 Iris dataset

The goal of this dataset is to predict the class of iris plant based on some physical characteristics of the petal and the sepal. The dataset counts with 150 instances, 50 instances each class (Iris Setosa, Iris Versicolour and Iris Virginica) and has 4 attributes (sepal and petal width and length). We pre-processed the data by one-hot-encoding the categorical classes and by replacing all the '?' values by NaN in order to remove them easily if they existed. This dataset showed very high correlation between features but we could not remove them for two reasons: first, because it doesn't have many features and second, because it makes sense for different physical attributes of a petal or sepal to be correlated.

4 Results

4.1 Experiments on Logistic Regression

In order to test the performance of our linear regression model, we run a set of experiments in which we varied the learning rate, the epsilon (or termination criteria) and the number of iterations for gradient descent. We observed that, for both abalone and iris datasets the accuracy of the model was unchanged by the value of epsilon picked; whereas for the adult and ionosphere datasets, the accuracy decreased when the value of epsilon was higher than 0.04.

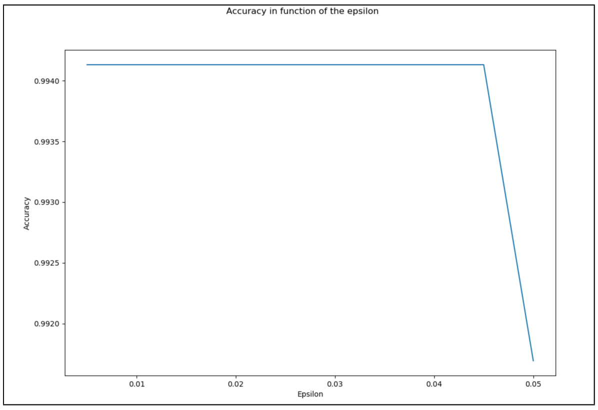

Figure 1: Adult

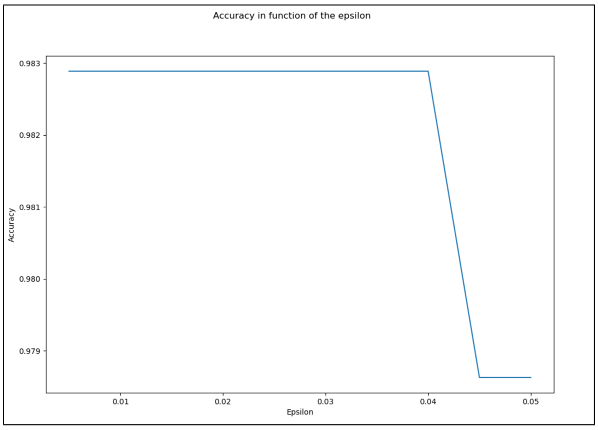

Figure 2: Ionosphere

Additionally, we observed that the accuracy increased with the number of iterations and the learning rate for all datasets. Here are some examples:

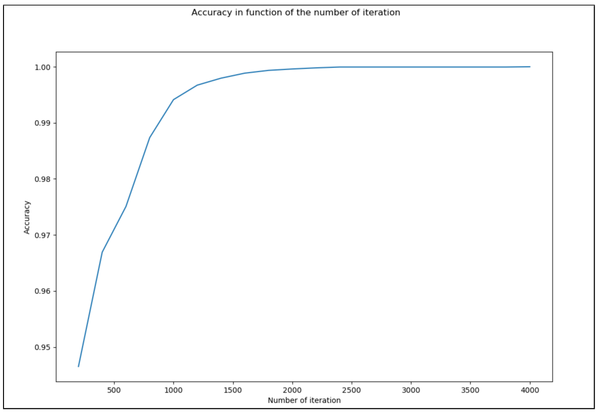

Figure 3: Adult

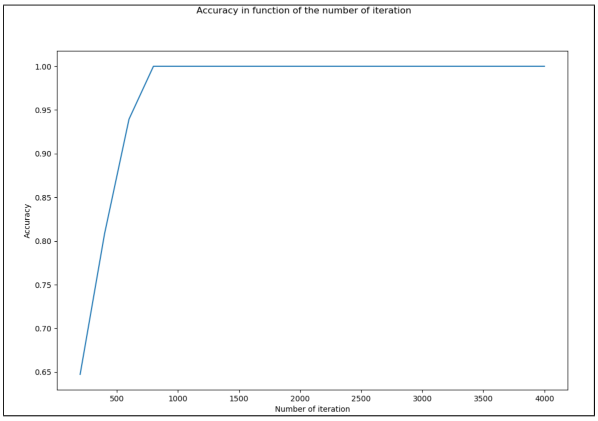

Figure 4: Iris

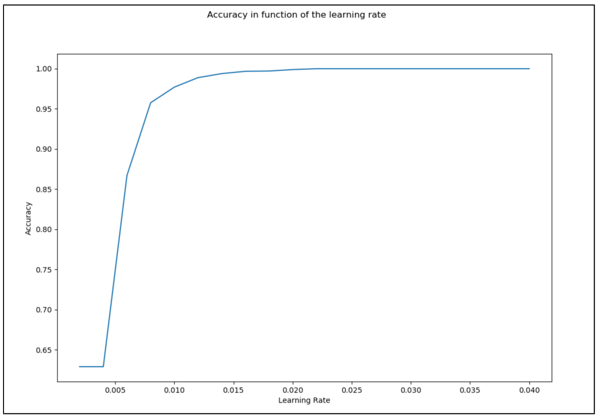

Figure 5: Abalone

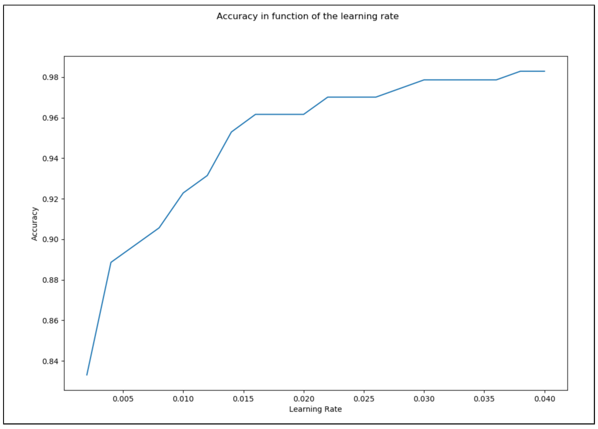

Figure 6: Ionosphere

4.2 Naive Bayes vs. Logistic Regression

Our next set of experiments was to compare the performance of both algorithms. We first compared the accuracy of the models for each dataset and observed that the Naive Bayes had a better performance for all of them (except for the training data of the iris dataset).

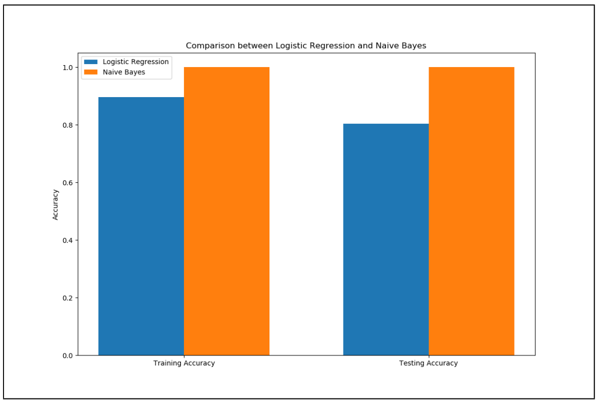

Figure 7: Ionosphere

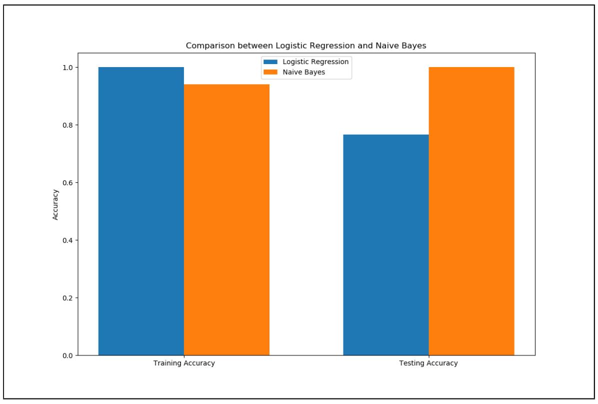

Figure 8: Iris

Finally, we tested the accuracy of both models as a function of the datasets' size. For the abalone dataset, we observed that in the logistic regression model, the train set accuracy was unchanged by its size meanwhile in the naive bayes model the greater the training set was the better was the accuracy. For the ionosphere dataset, the performance remained the same in the naive bayes model for all sizes of training set whereas in the logistic regression model the accuracy was increased with a larger training set. In the iris case, the accuracy remained similar in both naive bayes and logistic regression models irrespective of the training set's size; and finally, for the adult dataset, the performance was improved in naive bayes with an increase in the training set's size but remained the same in the case of logistic regression.

Regarding the features selection, we ran a few tests where we added other features, but we didn't observe any significant improvement while the running time became longer.

As for the parameters selection, we implemented a method find_parameters that optimizes the learning rate, the epsilon threshold and the maximum number of iterations in order to find a good subset of parameters to run the models.

5 Conclusion

Overall, our Naive Bayes model seems to perform better on most of the tests we implemented. However, we observe that for the iris dataset, logistic regression works better during training. Since the iris dataset was the only one we tested containing more than two it might be true that naive bayes works better on binary classification whereas logistic regression is more suitable for multiclass tasks. We would need to test our models on other multiclass datasets such as the glass identification dataset [2] or even the abalone dataset without our tweak to confirm this.

Still, both of our models yield pretty good result, which is promising. We would need to realise more tests in order to have a better evaluation of our models.

6 Statement of contribution

We all contributed in all sections by reviewing / testing each other's code.

- Mamoun Benchekroum: Task 1 + Plots
- Ines Moreno: Task 2 (Linear Regression) + Write-Up
- Julien Mounthanyvong: Task 2 (Naive Bayes) + Task 3

References

[1] A. S. Chatterjee S. Hadi and B Price. *Regression Analysis by Example (Third ed.)* John Wiley and Sons., 2000.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository.* 2017. URL: `http://archive.ics.uci.edu/ml`.

[3] Richard Goldstein and David Belsley. "Conditioning Diagnostics: Collinearity and Weak Data in Regression". In: *Technometrics* 35 (Feb. 1993), p. 85. DOI: `10.2307/1269293`.

[4] Andrew Ng and Michael Jordan. "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes". In: *Adv. Neural Inf. Process. Sys* 2 (Apr. 2002).

[5] V G Sigillito et al. "Classification of radar returns from the ionosphere using neural networks". In: *Johns Hopkins APL Tech. Dig* vol. 10 (1989). in, pp. 262–266.