Miniproject 1 - Machine Learning 101

COMP551 - Applied Machine Learning

Ege Odaci
McGill University
ege.odaci@mail.mcgill.ca

Rafael Gomes Braga
École de Thechnologie Supérieure
rafael.gomes-braga.1@ens.etsmtl.ca

Ramon Figueiredo Pessoa
École de Thechnologie Supérieure
ramon.figueiredo-pessoa.1@ens.etsmtl.ca

***Abstract*—In this mini-project we studied the performance of two classification models, namely Logistic Regression (LR) and Naive Bayes(NB), on four benchmark datasets. We analysed the features and labels of each dataset and applied preprocessing techniques to prepare the data to be used with the learning algorithms. We then performed some experiments to learn how the models behave in relation to each dataset. For both models we evaluated their performance with different sizes of training sets and for Logistic Regression we tried different values for the learning rate. After finding the best values for those parameters, we used 5-fold cross validation and choose the best performing models. We learned that both models performed comparably for all datasets, except for the Wine Quality one, in which Logistic Regression performed poorly. Finally, we compared our implementations to the corresponding ones provided by the scikit-learn package.**

I. Introduction

In this first project we experimented on the performance of the two significant classification models Logistic Regression and Naive Bayes on all four datasets. To begin with each dataset was analyzed to see what kind a information is stored in them. Ionosphere data was collected in Gose Bay, Labrador using 16 high frequency antennas and reported the signals and a label classifying these signals as b for bad and g for good. As can be seen from ionosphere.name file, even indexed columns are used to store the real values whereas the odd indexed columns are used to store complex values thus there is no correlation between them. First column consists of many 1's in contrast to second column being filled with many 0's. [1] Adult data involves some non continuous information such as working class, education, marital status, occupation, relationship, race, sex and native country recorded as strings. Adult data also involves some continuous information which are age, total education in years, capital gain, capital loss, hours per week and final weight recorded as integers [2]. At the very last column it specifies for each person whether they make more or less than 50K US dollars. Our third dataset is about breast cancer cells which collected from University of Wisconsin Hospitals. Each record have id number and some important attributes rated between 1 to 10. At the last column it specifies the cancer cell as 2 for Benign or 4 for Malignant [4]. Final dataset is about Wine Quality, It has scientific information stored as floats about the ingredients of the wine such as pH, fixed acidity etc. Last column stores the general rating for each wine, calculated by the median of at least 3 wine expert's rating [3]. Logistic Regression and Naive Bayes are tested on each dataset with different training sizes then their accuracies were compared to each other. For Ionosphere dataset both Logical Regression and Naive Bayes outputted nice accuracy with Logical Regression having slightly higher accuracy. For the adult dataset we observed that Logical Regression has provided better accuracy. With Wine dataset, Naive Bayes has given significantly high accuracy. For the breast cancer dataset, both models have really high accuracy with Logical Regression having higher accuracy.

II. Datasets

A. Output Labels

Since we are studying binary classification tasks, we needed to be sure that each dataset only outputs two possible labels. This is true for all datasets, except the Wine Quality one, which outputs a number from 1 to 10. We converted that dataset to a binary task by changing its output to 0 where the original value was less than or equal to 5 and 1 otherwise. We also computed the distributions of the two classes for each dataset and summarize it in Figure 1.

B. Missing Data and Malformed Features

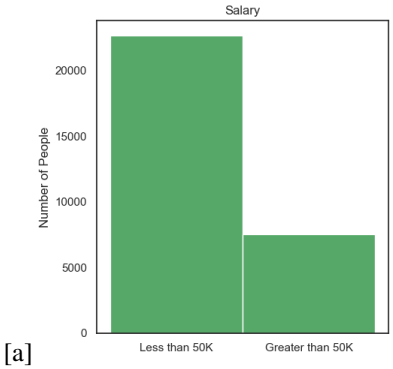
The Adult and Breast Cancer Diagnosis datasets have points with missing values for some of the features, represented by the '?' symbol. We decided to remove those points. We also found some malformed features and irrelevant features:

- In the Ionosphere dataset 89.2% of the values of the first feature and all the values of the second feature are zero
- In the Adult dataset the columns labeled "capital-gain" and "capital-loss" are also mostly composed of zeros (91.7% and 95.3% respectively)
- The first feature of the Breast Cancer dataset is the sample code number, which has no influence if the cancer is benign or malignant

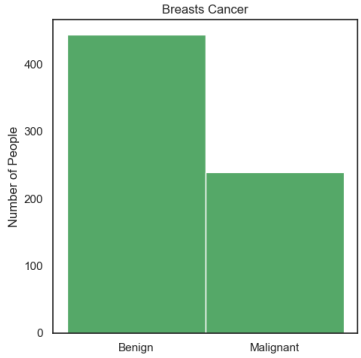
We decided to remove those features completely.

C. Continuous Fetures

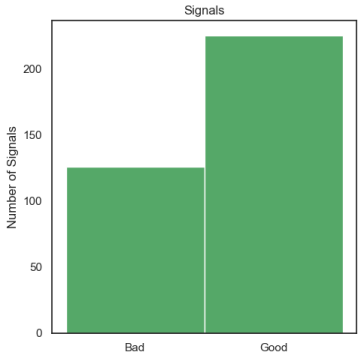
All features in the Ionosphere, Breast Cancer Diagnosis and Wine Quality datasets and 6 features in the Adult dataset are continuous. We plotted histograms and computed basic statistics for each of them in order to understand their distribution. We also applied stardardization to make the so the data has similar scales.



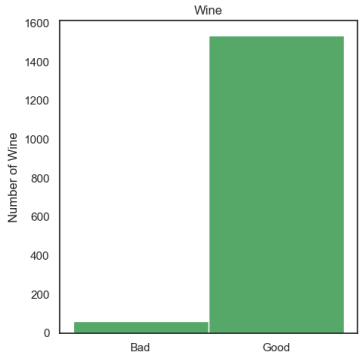
[a]



[b]



[c]



[d]

Fig. 1. Distributions of the positive vs. negative classes (histogram): a Adult; b Breast Cancer Diagnosis; c Ionosphere; and, d Wine Quality.

TABLE I
TESTING ACCURACY OF LINEAR REGRESSION (LR) AND NAIVE BAYES (NB) VERSUS THE CORRESPONDING SCIKIT-LEARN IMPLEMENTATIONS (SK-LR AND SK-NB) FOR EACH OF THE FOUR DATASETS

Dataset	LR	NB	SK-LR	SK-NB
Ionosphere	83.09%	85.91%	81.69%	76.05%
Adult	71.92%	72.66%	82.06%	75.21%
Wine Quality	56.25%	95.31%	96.56%	94.06%
Breast Cancer	96.35%	97.08%	96.35%	97.08%

Categorical Features

The Adult dataset have 8 categorical features. We applied one-hot encoding to convert those values to numeric.

D. Correlation between features

We draw heatmaps of feature versus feature to understand how they are correlated to each other. We found out that some of the features are highly correlated with other ones. Figure II-D shows some of the more interesting heatmaps.

III. RESULTS

A. Testing Different Training Sizes

In this experiment we trained our models against all datasets varying the size of the trainning data from 5% to 95%, and computed the accuracy of the prediction for each case. Figure 4 shows the results.

We observe that for some of the datasets the accuracy varies a lot as the function of the training size, while for others it is almost constant. We also tried shuffling the data before splitting it into training and testings sets. Figure **??** shows the results obtained after shuffling.

We observe that shuffling actually increased the overall accuracy of the models. Based on this result, we choose to use the value 80% for the splitting, with shuffling.

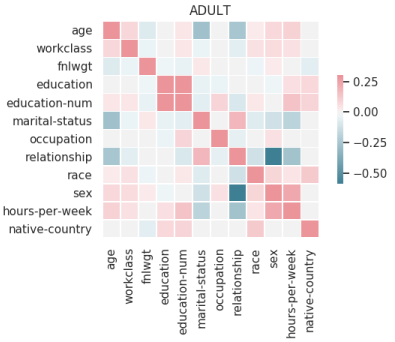
B. Different Values of Learning Rate for Logistic Regression

In this test we ran Logistic Regression against all datasets varying the learning rate to understand which value would give us the fastest training time. Image III-B shows a plot of the cost function against number of iterations for varying learning rates against the Adult dataset. The plots for the other datasets have a similar format.

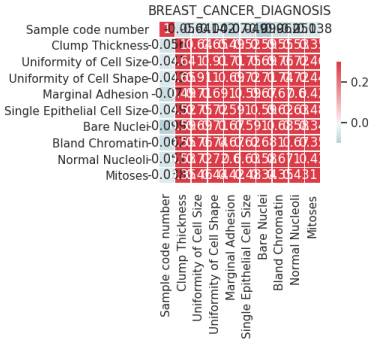
Here we noticed that, as expected, increasing the learning rate increases the speed of the training process. We decided to use the value of 1 for our next tests since it makes the cost function converge very fast.

C. Comparison between Logistic Regression and Naive Bayes

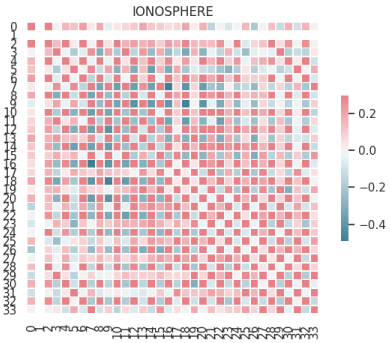
After selecting the best parameters we trained the Logistic Regression and Naive Bayes models against all datasets to compare their performances. We also wanted to compare our implementation to industry standard ones so we imported the scikit-learn and used their implementation of the two algorithms on the same datasets. Table I summarizes the results for all runs we did.



[a]



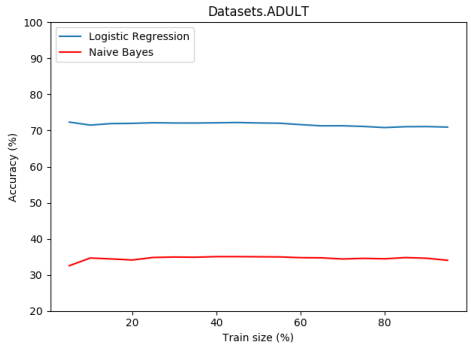
[b]



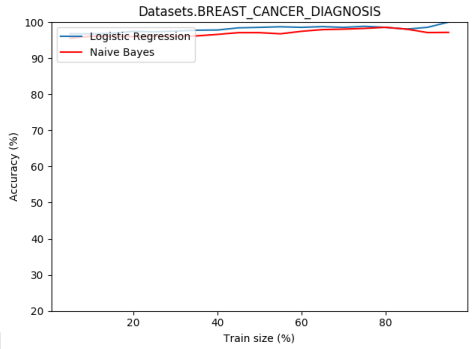
[c]

[d]

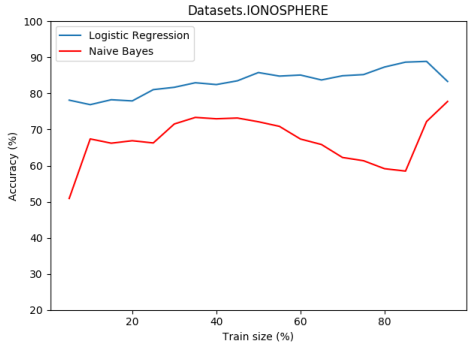
Fig. 2. Dataset heatmaps (correlation matrix): a Adult; b Breast Cancer Diagnosis; c Ionosphere; and, d Wine Quality.



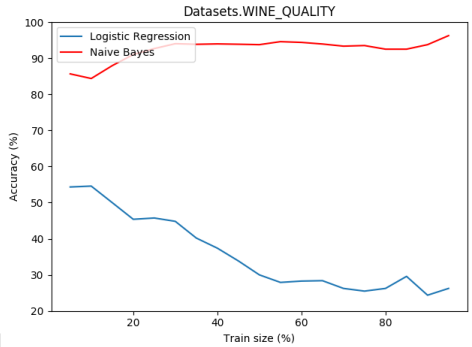
[a]



[b]



[c]



[d]

Fig. 3. Comparing the accuracy of the two models (LR and NB) as a function of the size of the training set): a Adult; b Breast Cancer Diagnosis; c Ionosphere; and, d Wine Quality.

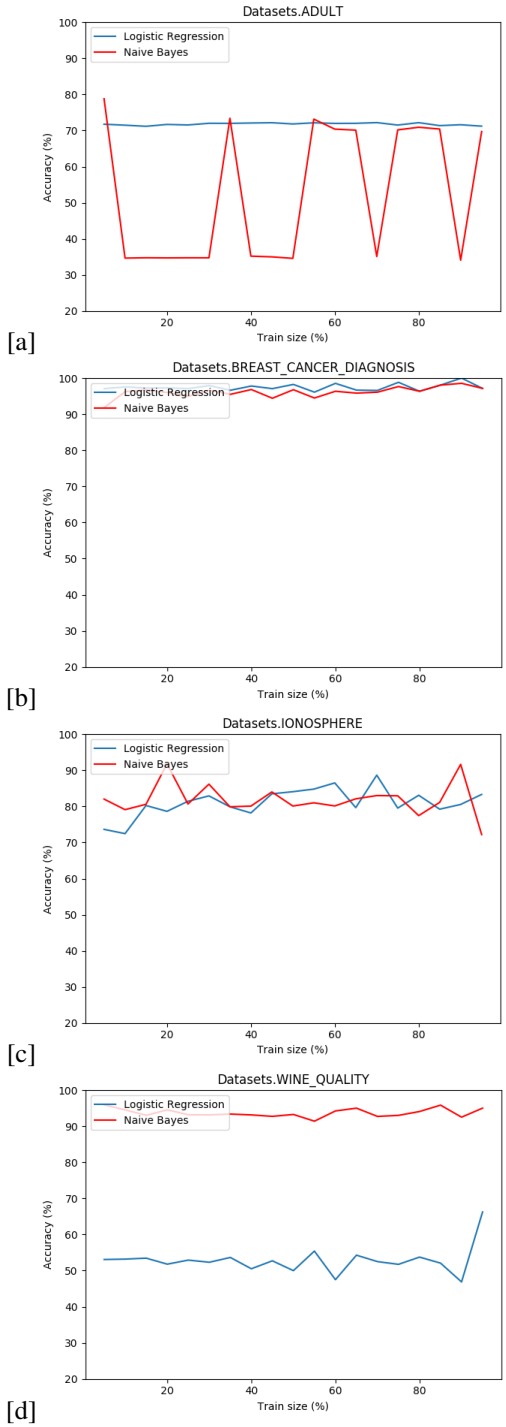


Fig. 4. Comparing the accuracy as a function of training size with shuffling: a Adult; b Breast Cancer Diagnosis; c Ionosphere; and, d Wine Quality.

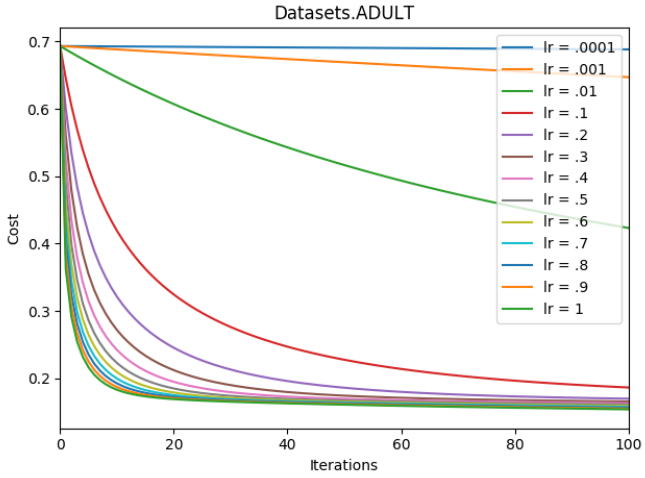


Fig. 5. Cost function versus number of iterations for Logistic Regression against the Adult dataset for different learning rates

TABLE II
K-FOLD CROSS VALIDATION (K=5) SCORES: OF LINEAR REGRESSION (LR) AND NAIVE BAYES (NB) VERSUS THE CORRESPONDING SCIKIT-LEARN IMPLEMENTATIONS (SK-LR AND SK-NB) FOR EACH OF THE FOUR DATASETS

Dataset	Scores (%): K-fold cross validation (k=5)
LR: Ionosphere	[80.35, 83.92, 82.14, 80.35, 85.71]
LR: Adult	[71.93, 71.82, 71.59, 71,47, 71.55]
LR: Wine	[48.04, 49.60, 50, 51.17, 57.64]
LR: Cancer	[94.54, 98.16, 97.24, 97.24, 98.16]
SK-LR: Ionosphere	[87.50, 87.50, 76.78, 87.50 , 87.50]
SK-LR: Adult	[81.88, 82.11, 81.85, 82.70, 82.57]
SK-LR: Wine	[96.48, 96.09, 95.70, 94.53, 97.25]
SK-LR: Cancer	[96.36, 97.24, 95.41, 95.41, 97.24]
NB: Ionosphere	[75.00, 92.85, 83.92, 83.92, 73.21]
NB: Adult	[36.30, 36.06, 37.77, 34.81, 37.00]
NB: Wine	[94.53, 92.18, 94.14, 93.35, 92.54]
NB: Cancer	[92.72, 97.24, 97.24, 95.41, 98.16]
SK-NB: Ionosphere	[82.14, 87.50, 73.21, 91.07, 75.00]
SK-NB: Adult	[69.81, 70.42, 70.76, 34.18, 71.27]
SK-NB: Wine	[92.57, 94.53, 96.87, 93.75, 92.94]
SK-NB: Cancer	[95.45, 96.33, 94.49, 96.33, 98.16]

Here we can see that our Naive Bayes implementation obtained results comparable to the scikit-learn implementation for all datasets. The Naive Bayes model, however, obtained bad results agains the Wine Quality dataset.

We also implemented and applied cross validation and summarize the results in Table II.

IV. Discussion and Conclusion

After performing all the experiments, we were able to improve our models obtaining good results against all datasets, except for one case: Logistic Regression against the Wine Quality dataset. Unfortunately we didn't have enough time to investigate further why that was happening. However, we used a similar implementation with the scikit-learn package to compare to algorithms and observed that the performances were comparable. We learned that one of the most important steps is the data analysis and preprocessing. For the future, it

would be interesting to work again with the Wine dataset using its output column original value, treating it as a multiclass problem.

V. Statement of Contributions

Ege worked on the visualization and analysis of the datasets, and implemented the *evaluate_acc* function. Rafael implemented the machine learning models and performed the experiments. Ramon worked on the preprocessing of data, provided the *sklearn* implementation that we used as benchmark, developed the k-fold cross validation script and created the command line interface.

References

[1] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker *"Classification of radar returns from the ionosphere using neural networks,"* Johns Hopkins APL Tech. Dig. Applied Phys. Lab., 1989.

[2] Ron Kohavi *"Scaling Up the Accuracy of Naive Bayes Classifier: a Decision Tree Hybrid"* Johns Hopkins APL Tech. Dig. Applied Phys. Lab., 1996.

[3] 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. ISSN: 0167-9236.

[4] .N. Street, W.H. Wolberg, O.L. Mangasarian. *"Nuclear feature extraction for breast tumor diagnosis.."* 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.