COMP 551: Applied Machine Learning
Mini-Project 1

Patrick Brebner 260889870, Mohamed Maoui, Dannie Fu 260874695

February 16, 2020

Abstract: In this project, we investigated and compared the performance of logistic regression using gradient descent for optimization and the Naive Bayes classification model on four different datasets. The accuracy of each model was tested by varying parameters, such as learning rate and dataset size, to determine how each one performs under different conditions. We found that although both models worked relatively well and had similar classification performance, there were many factors, such as feature selection, dataset size, and other hyper-parameters, that impacted these results.

1. Introduction

Logistic regression and Naive Bayes are two popular classification models for machine learning. Although similar, these two models function off of different underlying principles. Logistic regression is a discriminative linear classifier that learns the posterior probability and tries to find a decision boundary that best separates each class, whereas Naive Bayes is a generative model that learns the joint probability of the data and tries to determine the maximum likelihood that a class generated a certain instance. Naive Bayes works on the assumption that all the features are independent. This project investigates and compares these two classification models using four datasets: the ionosphere dataset, adult census dataset, the breast cancer diagnosis dataset, and the white wine quality dataset.

Previous analysis has been carried out on these datasets to explore different methods and approaches to classification. In Chockalingam et al.'s report, they compare various machine learning models such as Naive Bayes, Logistic Regression, Neural Networks, and Support Vector Machines on the Adult Census dataset. They suggested that logistic regression did not perform as well as other classifiers, such as the decision tree classifier which we have not covered in class, as it doesn't take into account the relationships between features, and that Naive Bayes also did not perform as well due to the conditional independence assumption. [1]. In the paper by Sigillito et al, they tested several feedforward neural networks on the Ionosphere dataset to investigate the effect of single layer vs multi layer network type affects performance. In this dataset, they took the first 200 instances as the training set to ensure that the good versus bad radar returns were equally balanced [2]. P.Cortez et al, investigated three regression techniques to predict human wine taste preferences. They found that Support Vector Machines outperformed multiple regression and neural network methods [3]. Finally, in the paper, "Classification and Diagnostic Prediction of Breast Cancers via Different", Saygla investigates six different classification methods, and uses a technique called Gain Ratio to do feature selection. After feature selection, they use 24 of the total 32 features [4].

Many papers have investigated classification techniques with these four datasets. In this report we also explored the performance of logistic regression and Naive Bayes on the datasets. We found that both models shared similar classification performance, however the logistic regression model worked better even with correlated features, whereas the Naive Bayes model had lower performance for datasets where the features were correlated. Furthermore, the Naive Bayes usually would converge faster and have higher accuracy for smaller datasets due to the prior probabilities.

2. Datasets

We analyzed four datasets in this project: the Ionosphere dataset, the Adult Census dataset, the Breast Cancer Diagnosis dataset, and the White Wine Quality dataset, all of which were obtained from the UCI Machine Learning Repository. Instances where there was a missing feature were removed from the dataset and features that did not contribute to the analysis or had skewed data were also removed.

2.1. Ionosphere Dataset

The Ionosphere dataset contains radar data and is used for classification of good versus bad radar returns from the ionosphere. The data was collected by a system in Goose Bay, Labrador, where there are 17 pulse numbers. There are 34 attributes in the dataset, 2 for each pulse, and all of which are continuous, plus the class attribute. Since there were no missing values in the dataset, no instances were removed. The second column was removed since all the values were 0. Consequently, we also removed column one as the feature was from the same pulse number as column two. Plotting the counts of good and bad radar returns, we can see that there was a bias towards good radar returns. In order to compensate for this bias, we take the first 200 instances for training, as is done by Sigilitto et al. in their paper "Classification of radar returns from the ionosphere using neural networks", where the number of good and bad were split almost 50/50.

2.2. Adult Census Dataset

The Adult Census dataset contains census data on 48,842 people worldwide and is used to predict whether the income of a person exceeds 50K a year. The attributes in this dataset are: Age, Workclass, fnlwgt, Education, Education-Num, Marital Status, Occupation, Relationship, Race, Sex, Capital Gain, Capital Loss, Hours per Week, Native Country, and Income.

The fnlwgt attribute, which represents the final weight, was removed because it was deemed to not be a useful feature. The capital-gain and capital-loss features were also removed as most of the entries were 0. The native-country attribute was removed because most observations were taken from the United States and therefore would have biased the results. Rows with missing data (shown by a "?" or NaN) were dropped as well. The race attribute was also dropped because most of the race observations were recorded as being "white".

The dataset was then split into two classes, where the positive class was comprised of instances where the person earned more than 50K annually and the negative class was comprised of instances where the person earned less than or equal to 50K annually. The distribution of data was plotted for these two classes for each attribute to obtain a better understanding.

The mean age of the positive class was 44 with a standard deviation of 10 years, and the mean age of the negative class was 37 with a standard deviation of 13 years. Those who earned more than 50K a year seem to be in the middle of their career, whereas those who earned less than 50K appear to be more early in their career. The majority of people who earned more than 50K a year also graduated with a Bachelor's degree or higher whereas in the negative class, the majority graduated from high school. As the level of education increased, there were a larger portion of instances that earned higher than 50K annually. In the positive class, the majority were married and worked in a managerial executive position or in a professional occupation, whereas for the negative class the majority were never married and worked in a clerical administrative position, a craft-repair occupation, or in some other service industry.

The next step of the processing was to convert the target feature into a binary output. This was done by one hot encoding the income attribute into greater than 50K income and less than 50K income. Then, since the classification goal was to predict whether someone earns more than 50K annually, the latter column was dropped. The other categorical features, workclass, education , marital-status, occupation, relationship , and sex, were also one hot encoded to transform them into a binary output for the logistic regression model. However, for the Naive Bayes model, we left the categorical features because we used the multinomial distribution for the probabilities. The categorical features for the multinomial Naive Bayes were first transformed into integer labels, from their original string labels.

2.3. Breast Cancer Diagnosis Dataset

The original Breast Cancer Diagnosis dataset contains clinical data from 699 cases that can be used to determine if the diagnosis is benign or malignant, labeled 2 and 4 in the target data column, respectively. Out of the 699 cases, about 34 percent were diagnosed as malignant. On top of the target data, the dataset included the following 10 features: ID Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. Each feature has a categorical ranking from 1 - 10.

For processing this dataset, the first thing was to completely remove the ID Number feature as this would not help with the classification models. The target data was also adjusted so benign results were 0 and malignant results were 1. finally, rows with missing data (shown by a "?") were dropped as well bringing the final input dataset to 683 cases and 9 features.

As far as correlation between features, Uniformity of Cell Shape and Uniformity of Cell Size were highly correlated. For positive (malignant) results the Clump Thickness feature had a low correlation with all other features and for negative (benign) results the Mitoses feature had very low correlation with the other features. For all features the distributions between positive and negative results show that negative results tend to have features with lower rankings, typically below 4, while the positive results are more spread out but include higher rankings all the way to 10.

2.4. White Wine Quality Dataset

The White Wine Quality dataset contains 4898 samples of the "Vinho Verde" wine, from the north of Portugal. The dataset includes 11 continuous features: Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, and Alcohol. The target class ranks the wine from 1 - 10 based on sensory data.

The first step on processing this data was to convert the target feature to be a binary output. This was done by setting any wine with a quality of 6 or higher, considered good wine, to 1 and all other wine to 0 to illustrate bad wine. With this limit, 66 percent of the wines were considered good. The remaining features were normalized to help with classification.

As far as correlation there wasn't much of a trend with entire features. The only correlations that stood out were a high correlation between Density and Residual Sugar, a high correlation between Free Sulfur Dioxide and Total Sulfur Dioxide, a low correlation between pH and Fixed Acidity, and finally a low correlation between Alcohol and Density, Residual Sugar, and Total Sulfur Dioxide. All of these correlations were not surprising. The distributions between good and bad results were very similar for all features, often looking like Gaussian distributions, with the one exception being alcohol content. The good wine's distribution was shifted to include higher alcohol amounts although only by a few percent.

3. Results

The results of the Logistic Regression and Naive Bayes models on each dataset will be discussed in the following section. The main focus will be on the accuracy of each model over varying hyper-parameters and how the performance of the models compare for different training dataset sizes. Initial five fold cross validation tests, with no tuning or adjustment gave the results seen in table 1.

Initial Results				
	Logistic Regression		Naive Bayes	
DataSet	Accuracy (%)	Recall (%)	Accuracy (%)	Recall (%)
Ionosphere	83	96.7	82	79.8
Adult Census	81.5	45.2	79.3	76.3
Wine Quality	75	87.8	70	77.4
Breast Cancer	96	93.8	88.8	70.4

Table 1: Initial Cross Validation Results of Logistic Regression and Naive Bayes

3.1. Logistic Regression Analysis

The Logistic Regression Model performance was analyzed for varying hyper-parameters related to gradient descent. The main parameters being adjusted were learning rate, termination threshold, and the number of iterations of gradient descent. All the following logistic regression experiments were performed using five fold cross validation to use an average accuracy.

3.1.1. Learning Rate

Figure 1 shows the impact of learning rate on the accuracy of logistic regression for the four datasets. For this test, we used learning rates 0.1, 0.01, 0.005, 0.001, 0.0005, and 0.0001. As expected, when the learning rate decreases, the speed of convergence decreases because the local minimum is approached in smaller increments. However, for small learning rates, accuracy can improve as there is less of a chance to miss or overshoot the minimum. For the large learning rates, the accuracy can be negatively effected because the model might miss the minimum or diverge, as previously mentioned. This negative effect likely led to the lower accuracy seen in the Adult Census dataset at the larger learning rates. Overall, the experiment shows that there is typically an ideal learning rate for the model/dataset but if made too small the model takes too long to converge and thus reduces the accuracy. This drop off in accuracy for very small learning rates can be clearly seen in figure 1.

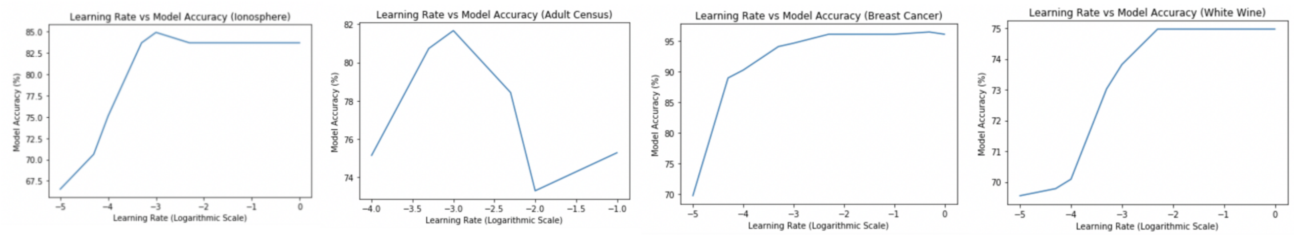

Figure 1: The effect of learning rate on logistic regression performance.

3.1.2. Termination Threshold

This experiment was performed to explore the effect of an early termination threshold, other than number of iterations, on gradient descent. The early termination criteria was based on the the size of the gradient and would terminate the gradient descent early if a minimum threshold was passed. Early termination is often used as a method to reduce over-fitting. Figure 2 shows the impact of the termination threshold on the accuracy of logistic regression. In general the accuracy decreased as the termination threshold was less strict. This makes sense as you would sacrifice accuracy in your cross validation to create a more general and overall more useful model for any future data.

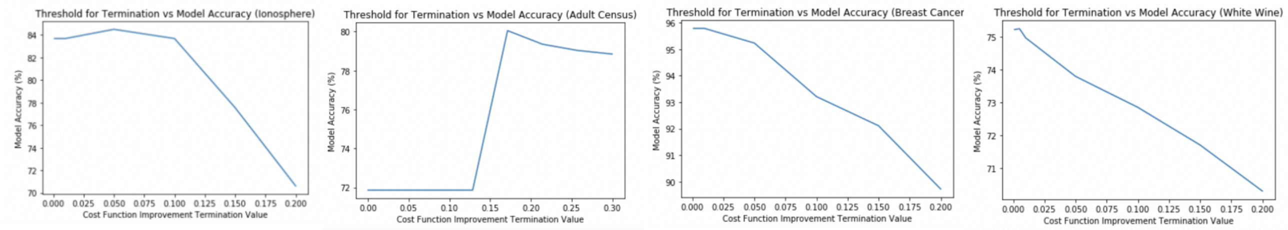

Figure 2: The effect of termination threshold on logistic regression performance.

3.1.3. Number of Iterations of Gradient Descent

In this experiment we wanted to evaluate the necessary number of gradient descent iteration required to reach maximum accuracy. Figure 3 shows the impact of the number of iterations in gradient descent on the accuracy of logistic regression. We tested from 100 iterations to 20000 iterations of gradient descent on three of the four datasets while running the Adult Census dataset up to 45,000 iterations. As expected, there was a sharp increase in accuracy from a low number of iterations which then leveled out in three of the four datasets, after which there was not much of an improvement. This means after a certain number of iterations there is really no point in increasing the number further and it would just be a waste of computational power. The Adult dataset seemed to keep improving with the number of iterations but with such a large and complex dataset the computational requirements for running this many iterations was quite demanding. In this case a compromise between accuracy and computational requirements may be required, depending on the hardware being used.

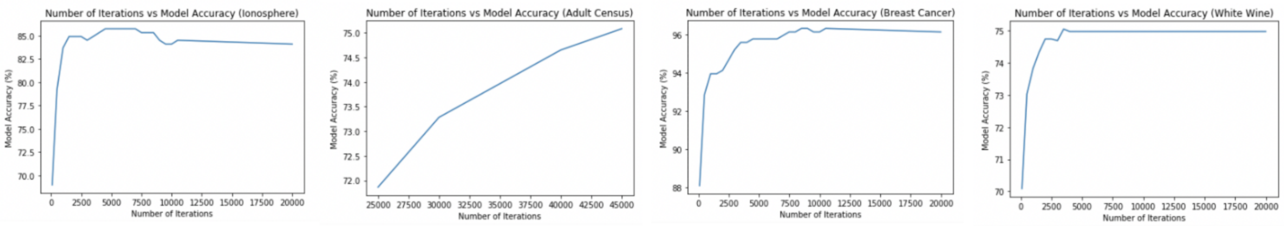

Figure 3: The effect of number of iterations of gradient descent on logistic regression performance.

3.2. Logistic Regression vs Naive Bayes

The performance of the Logistic Regression model was compared with the performance of the Naive Bayes model. We tested how the size of the dataset impacted the classification performance. Different probability distributions for the Naive Bayes implementation were used for the four datasets, as the type of features varied between datasets. For the ionosphere dataset and the white wine dataset, all the features were continuous so the

Gaussian Naive Bayes implementation was used. The adult census dataset and breast cancer diagnosis dataset had categorical features so Multinomial Naive Bayes was used.

3.2.1. Training Set Size and Model Accuracy

In this experiment, we wanted to investigate how the dataset size impacted the performance of the models. We looked at the accuracy of the models for 10 to 100% of the original dataset size, in increments of 10%. Figure 4 shows the impact of dataset size on the model accuracy using the testing sets. From the figure we can see that in general, with a smaller training dataset size the model accuracy is lower. We can attribute this to fact that the model is only shown a small portion of examples of the whole dataset. This appears to be especially important in Naive Bayes, as the classification of testing data depends on the distributions of the training set data; therefore, for a smaller training set, the distributions and likelihoods could be skewed.

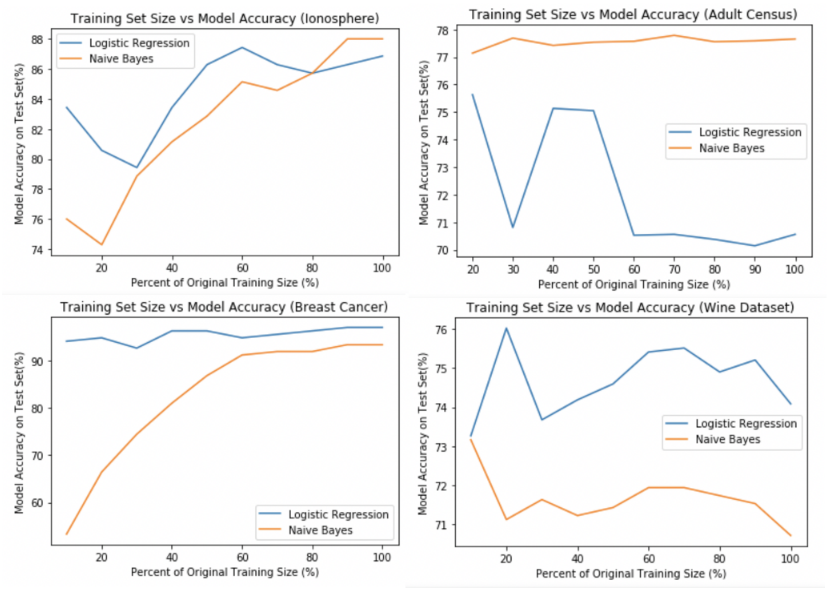

Figure 4: The effect of training set size on model performance.

4. Discussion and Conclusions

In this project the performance of two popular machine learning models, logistic regression and Naive Bayes, were compared over four datasets. Both models are used for classification but each have different strengths and weaknesses, some of which were explored through our experiments. The logistic regression model requires some hyper-parameter tuning in order to get the best accuracy but works reasonably well even with correlated features. The naive bayes model usually has decent accuracy for small datasets, due to prior probabilities, but since it assumes the features are independent the accuracy can be poor if there are dependent features. For example, in the adult dataset, the two features "Marital Status " and "Relationship Status" provide very similar information, so one of them could be removed. Future investigation into which other features of the datasets are dependent would allow us to remove unnecessary features and improve the accuracy of Naive Bayes. Another technique that could improve the Naive Bayes model would be Laplace smoothing to which would help for situations with zero likelihood. In conclusion, both models had similar performance but even with these relatively simple datasets it can be seen that the choice of model in machine learning is not arbitrary and that pre-processing of datasets is a crucial step in achieving ideal results.

5. Statement of Contributions

Patrick wrote the logistic regression model and the K fold cross validation script. Patrick also did preprocessing on datasets 3 and 4. Dannie did preprocessing on datasets 1 and 2 and wrote the multinomial naive bayes model. Mohamed wrote the Gaussian Naive bayes model.

References

[1] N. Chakrabarty, S. B. Biswas, Income classification using adult census data, Conference: 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)doi:10.1109/ICACCCN.2018.8748528.

[2] V. G. Sigillito, S. Wing, L. V. Hutton, K. Baker, Classification of radar returns from the ionosphere using neural networks, 1989.

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

[4] A. Saygılı, Classification and diagnostic prediction of breast cancers via different classifiers, International Scientific and Vocational Studies Journal 2 (2) (2018) 48–56.