Machine Learning for Engineers (ECSE 511)
Mini-Project 1

Jiarui Xie **Chonghui Zhang** **Siyuan Sun**

Abstract

In this report, logistic regression, one of the supervised learning models, is used to correlate between the features and targets. And machine learning technology is utilized to predict unknown targets with new input observations. The objective of supervised learning is to discover the pattern of datasets that can predict the outcome correctly. In these two datasets of Hepatitis and Bankruptcy, different types of features with a target attribute are given. In the dataset of Hepatitis, the basic features are used to build the logistic regression model with different learning rates and stopping criteria to test the performance of the model. Also, the impact of the stopping criterion and learning rate on the predictive accuracy and training runtime are investigated. 10-fold cross-validation is used to compare the models and select features. The subset of features and the combination of polynomial features that offer the highest accuracy through cross-validation are chosen. In this project, the performance of linear classification on two datasets is studied, the logistic regression approach achieved better accuracy when some features are removed. Python 3.6 is utilized to code the machine learning modules such as logistic regression, performance evaluation and cross-validation

1 Introduction

The objective of this mini project is to implement logistic regression with a linear classifier for the datasets of Hepatitis and Bankruptcy. Logistic regression is applies to discrete target variables such as binary responses. The goal is to build a binary classification model that predicts the outcome based on the input features. These two datasets have different natures and types of features, but both consist of observation (input features) and class label (output). The dataset of Hepatitis has both categorical and numerical features, while the dataset of Bankruptcy has only numerical features. Both datasets have binary target indicating yes/no class label. In the beginning, the datasets are preprocessed in order to perform data analysis and speed up the convergence. Then, the basic features of the two datasets are used to build the initial logistic regression models, with the entire datasets as the training sets to quickly observe the predictive performance. After, some features are added, and cross-validation is utilized to choose the best model. The impact of learning rate and stopping criterion on accuracy and runtime are investigated as well. For the Hepatitis dataset, the initial logistic regression model converged after 130 iterations with the average weight loss reached 0.03 and 65, respectively. After comparing different stopping criteria, the mean weight threshold 0.03 provides a better performance than maximum weight and loss function threshold. The learning rate of 0.27 could make the model achieve the shortest runtime. By using feature selection based on the Maximal Information Coefficient (MIC), decreasing the number of features to 2 leads to an increase of accuracy in the validation set from 77.3% to 84.9%. For the initial model of the Bankruptcy dataset, the maximum iteration number was set to be 20000 and gradient descent will be stopped when the average weight change becomes less than 0.2. Feature selection was needed for the improvement of the model since the model could only achieve the accuracy of 70.86%. By using feature selection based on MIC, decreasing the number of features to 3 lead to an increase of accuracy in the validation set from 65.3% to 70.7%.

2 Datasets

2.1 Hepatitis dataset

The main objective of the classification model built for the Hepatitis dataset is to differentiate two classes: survivors and patients whose hepatitis proved terminal. The dataset is comprised of different types of features for hepatitis patients, such as the patient's age, a numerical feature, and sex, a categorical feature. The dataset comes with a table of data with 19 feature columns and 1 output (target) column. 142 observations are provided. The target column is named as 'Class Label' indicating survivors or patients whose hepatitis proved terminal. The histogram shape of feature age can be seen as right-skewed distribution. And the histogram shape of feature protime shows the double-peaked distribution. Some plots of features are shown in Figure 1 and have been used to observe the potential dependencies. More distributions plots can be found in Appendix Figure 1 and 2.

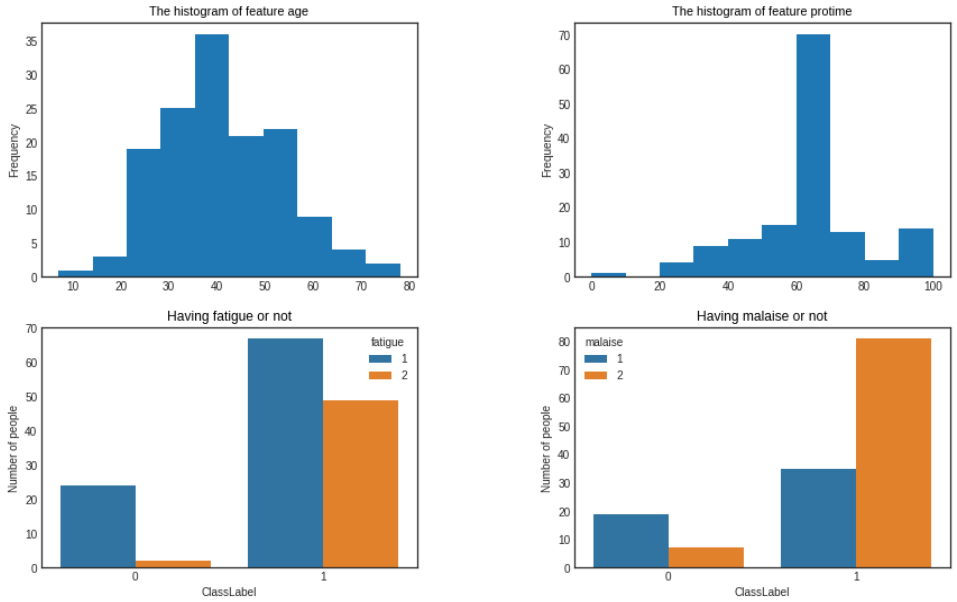

Figure 1: **The plots of some features of Hepatitis dataset**

2.2 Bankruptcy dataset

The Bankruptcy dataset consists of different econometric attributes for bankruptcy status. The dataset is comprised of 64 feature columns and 1 target column. 453 observations are provided. The target shown as 'ClassLabel' indicates whether the bankruptcy happens or not. After dropping some features, the histogram shapes of some left features are shown in Figure 2.

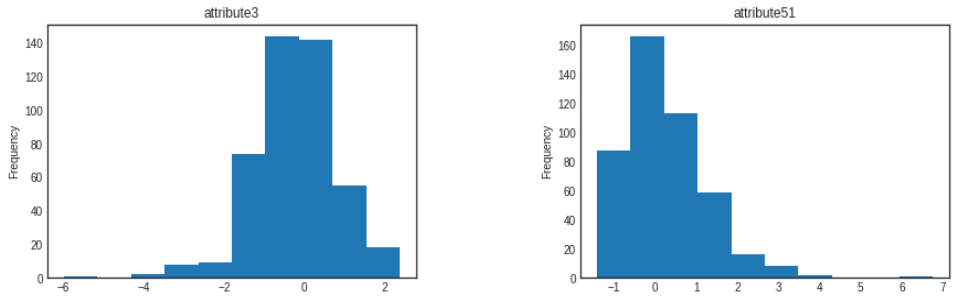

Figure 2: **The plots of some features of Bankruptcy dataset**

2.3 New features

In this project, we use a Polynomial Curve Fitting method to make our model able to predict the non-linear relationship. Hence, the function 'add_feature (olddata, order)' was built such that the basic features were transformed to polynomial features, and added to the original dataset. The order of polynomial can be manually set by the users. For example, the original features (olddata) are shown in Table 1. By using 'add_feature (olddata,2)', the output is the original features companied with the square of them, shown in Table 2.

Observation 1	X1,1	X1,2	Y1
Observation 2	X2,1	X2,2	Y2

Table 1: **The old data**

Observation 1	X1,1	X1,2	X1,1^2	X1,2^2	Y1
Observation 2	X2,1	X2,2	X2,1^2	X2,2^2	Y2

Table 2: **The new output**

2.4 Subset of features

Many machine learning models, especially those based on regression slopes and intercepts, will estimate parameters for every input feature. Because of this, the presence of noninformative variables can add uncertainty to the predictions and reduce the overall effectiveness of the model [1]. Therefore, it is also necessary to find the most suitable subset of features, which contains less non-informative variables. To achieve that, the Maximal Information Coefficient (MIC) is used to rank the variables and select the useful features. Feature selection using variance is also investigated

3 Results

For both datasets, the features and output (class label) were first segregated. Then, a dummy feature was added to the original feature to simplify the notation of the linear model. Then the numerical features were normalized while the dummy feature and categorical features were left unnormalized. To find the best weight W, the least-squares solution method was used as the loss function to calculate the error between the training data and predicted results. And gradient descent was utilized to locate the global minimum of the loss function, which results in the minimum difference between the training data and predictions [2]. The starting location of gradient descent was randomized each time the logistic regression module is executed to avoid local minima [3].

A suitable stopping criterion reduces training runtime and guarantees high predictive performance [4] [5]. Gradient descent is usually stopped when the weights or loss function changes slowly. Three methods are investigated in this report: mean weight change, maximum weight change and loss function change. If the aforementioned values compared to the previous steps vary steadily within a threshold, the training process is terminated. To determine the best threshold for each method so that three methods can be compared, the training module built for Hepatitis was run repeatedly with the thresholds gradually increasing. The maximum value that guarantees a highest F1 score achievable is the best threshold for each method, which was found to be 0.03, 0.1 and 0.01 for mean weight, maximum weight and loss function, respectively. The effect of the learning rate will also be studied, and the range is from 0.03 to 0.45, which included learning rates of generally low and high values [6].

To improve the predictive performance of both models, feature selection will be performed with MIC and polynomial features. To find the best combination of features, 10-fold cross-validation will be used to calculate the accuracy of the training and validation sets for comparison.

3.1 Hepatitis dataset

Initially, the basic features were fed into the logistic regression model with a learning rate of 0.06 to quickly test the performance. The maximum iteration number was set to 5000 and gradient descend will stop when the average weight of the current step changed less than 0.03 compared to the weight from the last step. As shown in Figure 3a and 3b, the model quickly converged after 130 iterations, when the average weight loss function reached 0.03 and 65, respectively. An accuracy of 88.3% and F1 score of 0.892 were obtained, which indicated a satisfactory performance with both high precision and recall. Then, this model will be used to investigate the impact of stopping criteria and learning rate in terms of predictive accuracy and training runtime.

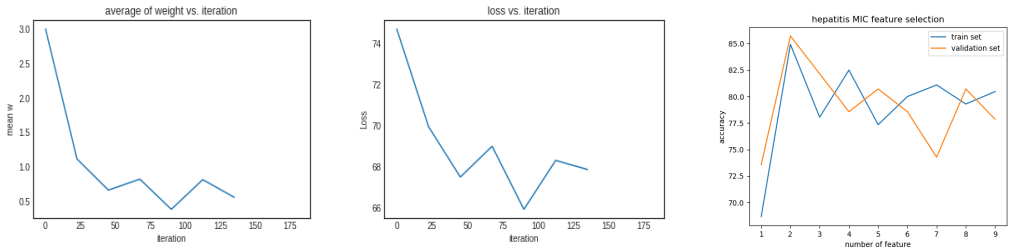

Figure 3: **(a) average weight vs. iteration and (b) loss vs. iteration (c)Accuracy of different orders of Hepatitis dataset**

The performances of three stopping criteria are rendered in Appendix Table 1. By comparison, the most suitable stopping criterion is the one that offers the shortest runtime, since the highest accuracy achievable is the same. Five tests were conducted for each criterion, as the starting weights are randomized. It could be observed that while achieving the same average F1 score, loss function as the threshold offered lower accuracy than other methods. Moreover, mean weight as the threshold provided a shorter runtime and a smaller number of iterations to yield the same level of accuracy and F1 score than max weight. Therefore, the most suitable stopping criterion is the mean weight for this question.

The effect of the learning rate is demonstrated by training a model with varying learning rates, but the same basic features and stopping criteria, which is a mean weight threshold of 0.03. The range of learning rate was from 0.03 to 0.45, taking a spacing of 0.03. Because the starting weights are randomized, there could be some variation of runtime at the same learning rate. Thus, the average iteration, runtime and evaluation metrics were recorded from 5 tests at each learning rate (Appendix Table 2). As figure 4 indicated, the iterations before convergence gradually decreased from 0.03 to 0.27. Then, there showed more fluctuation and an increasing trend after 0.27. Hence, 0.27 is the optimum learning rate with respect to the shortest runtime.

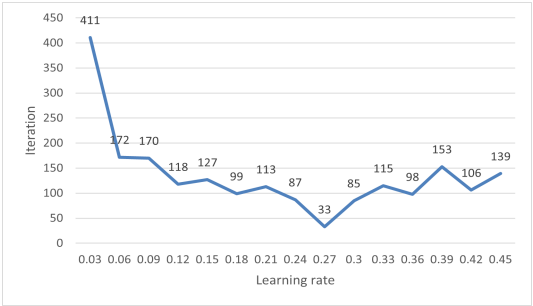

Figure 4: **The iterations required to converge to the threshold at different learning rates**

Figure 3c shows the relationship between the accuracy and number of features selected when using MIC. The training accuracy and validation accuracy were acquired with cross-validation. Each feature in Hepatitis dataset was scored and ranked based on MIC. Then, the logistic regression module was run repeatedly, and each time a feature with less score was removed. When the top two features were retained, the model is 84.9% accurate for the training set and 85.7% accuracy for the validation

set, which is the peak accuracy. That means the first 2 features, bilirubin and albumin, contains the majority of the useful information. Therefore, Polynomial Curve Fitting method will be used with the first 2 features. However, after the test, Polynomial Curve Fitting method did not provide a better accuracy.

3.2 Bankruptcy dataset

The basic features of Bankruptcy dataset were input into the logistic regression model with a learning rate of 0.1 to quickly test the performance. The maximum iteration number was set to 20000 and gradient descend will stop when the average weight of the current step changed less than 0.05 compared to the weight from the last step. As shown in the figure 5a and 5b, the model stopped at the maximum iteration iterations, when the loss function reached around 303. The first model was 70.86% accurate, which indicate that this model demanded feature selection for improvement.

Figure 5c shows the relationship between the accuracy and number of features left when feature selection based on MIC is used. When the number of features is 2, the model has got 69.1% accurate in the training set and 68.7% accuracy in the validation set. The peak accuracy is found when the number of features left is 3, 70.6% accuracy in training set and 70.7% accuracy in validation set. That means the first 3 features, attribute3, attribute8 and attribute34, contains majority of the useful information. Therefore, Polynomial Curve Fitting method will be used on first 3 features. However, after test, accuracy just fluctuated around 70% in first 4 orders, then dropped.

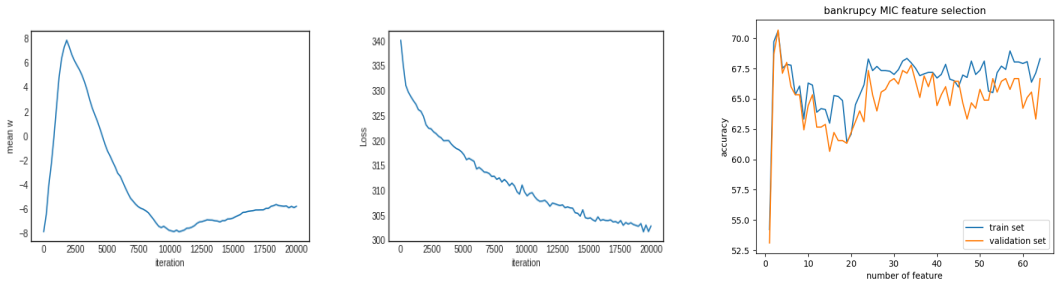

Figure 5: **(a) average weight vs. iteration and (b) loss vs. iteration (c)Accuracy of different order of Bankruptcy dataset**

Figure 6 shows the relationship between the accuracy and threshold of variances when the method of removing features with low variance is used. For both cases, Bankruptcy and Hepatitis, this method did not have good performances.

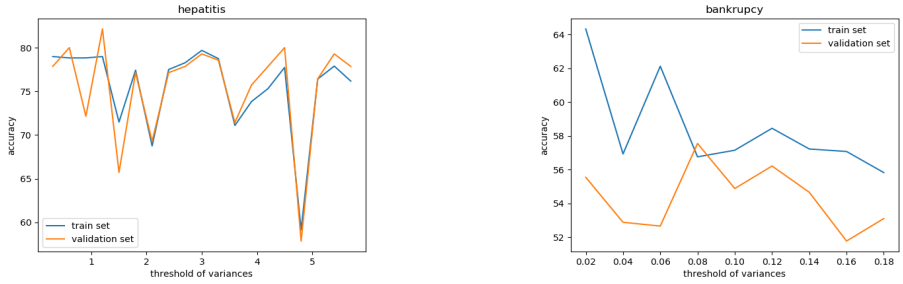

Figure 6: **Accuracy vs. threshold of variances of (a)Heptitis and (b)Bankruptcy dataset**

4 Discussion and Conclusion

When loss function is the threshold, the F1 score is greater than accuracy. By inspection, these models had a high recall but low precision, indicating that many negative observations were predicted to be

positive. Thus, using loss function resulted in a model that was biased to positive predictions. Maximum weight as the threshold posed a strict requirement of convergence; and thus, the training runtime was significantly greater than mean weight to offer the same predictive performance. Admittedly, only three stopping criteria were investigated so that it's not enough to conclude that mean weight is the best criterion. However, this report provides a framework for scholars to compare between different stopping criteria. As the learning rate increasing, gradient descent took bigger steps at each iteration. And the runtime reached the minimum when it decreased to 0.27. However, as the learning rate further increases, the steps tend to exceed the global minimum of the loss function, then move back-and-forth for many iterations until it lands within the threshold by chance. It could be seen in Figure 4 that after 0.27, the convergence became slower and slower with high fluctuation.

For feature selection, in both Bankruptcy and Hepatitis datasets, removing features with low variance did not improve the performance. But feature selection based on the Maximal Information Coefficient (MIC) improved the performances of models. In Hepatitis, by decreasing the number of features to 2, the accuracy in validation set rises from 77.3% to 84.9%. In Bankruptcy, by decreasing the number of features to 3, the accuracy in validation set rise from 65.3% to 70.7%. The improvement is due to the elimination of non-informative variables, which reduced the uncertainty of the model. For adding features, Polynomial Curve Fitting method is used in this project. However in both case, the improvement of performance is not obvious.

5 Statement of Contributions

Jiarui Xie was responsible for the construction of logistic regression classes, investigation of stopping criteria and tuning of learning rate.

Chonghui Zhang was responsible for the construction of k fold cross validation classes, investigation of adding new features and subset of features.

Siyuan Sun was responsible for the observation and statistical analysis of the dataset and the visualization of two datasets.

References

[1] Kuhn, Max, and Kjell Johnson. Applied predictive modeling. Vol. 26. New York: Springer, 2013.

[2] T. Hastie, J. Friedman, and R. Tisbshirani, The Elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2017.

[3] M. Swamynathan, Mastering Machine Learning with Python in Six Steps. Bangalore, Karnataka: Apress, 2017.

[4] S. Baghal, C. Paquette, and S. Vavasis, "A termination criterion for stochastic gradient descent for binary classification," arXiv, 2003.

[5] V. Patel, "Stopping Criteria for, and Strong Convergence of, Stochastic Gradient Descent on Bottou-Curtis-Nocedal Functions," arXiv.org, 2004. [Online]. Available: https://arxiv.org/abs/2004.00475. [Accessed: 11-Oct-2020].

[6] P. Surmenok, "Estimating an Optimal Learning Rate For a Deep Neural Network," Medium, 14-Jul-2018. [Online]. Available: https://towardsdatascience.com/estimating-optimal-learning-rate-for-a-deep-neural-network-ce32f2556ce0. [Accessed: 11-Oct-2020].

Appendix

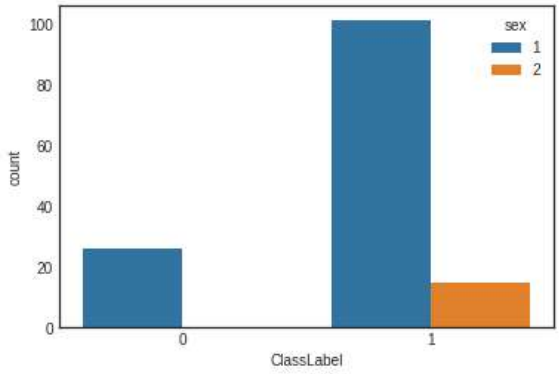

Figure 1. The count plot of feature sex

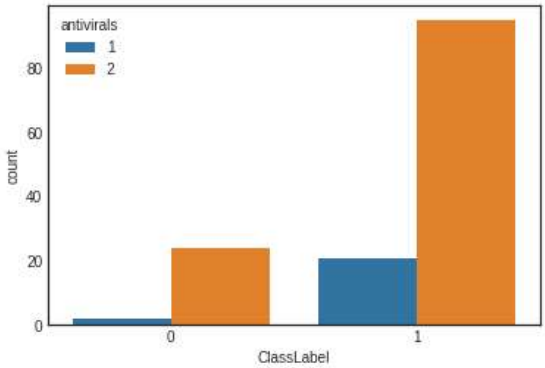

Figure 2. The count plot of the feature antivirals

Stopping Criteria	Iteration	Runtime (t)	Accuracy (%)	F1 score
mean weight Threshold = 0.03	89	0.009	88.732	0.896
	79	0.009	88.028	0.895
	36	0.004	86.620	0.895
	64	0.007	86.620	0.894
	159	0.016	88.732	0.896
max weight Threshold = 0.1	683	0.062	88.732	0.897
	637	0.053	88.732	0.897
	603	0.053	88.732	0.897
	798	0.064	88.732	0.897
	1063	0.091	90.141	0.897
loss function Threshold = 0.1	60	0.001	81.690	0.903
	298	0.038	81.690	0.903
	42	0.004	76.761	0.877
	40	0.004	85.915	0.900
	44	0.004	83.099	0.903

Table 1. Comparison among the performances of three stopping criteria.

learning rate	Average iteration	Runtime(t)	Accuracy (%)	F1 score
0.03	411	0.064	88.732	0.897
0.06	172	0.020	88.028	0.896
0.09	170	0.020	89.437	0.897
0.12	118	0.012	88.732	0.896
0.15	127	0.013	88.732	0.896
0.18	99	0.010	88.732	0.896
0.21	113	0.012	88.028	0.896
0.24	87	0.009	88.732	0.896
0.27	33	0.004	88.915	0.895
0.3	85	0.009	88.732	0.896
0.33	115	0.012	88.732	0.896
0.36	98	0.010	88.732	0.896
0.39	153	0.016	88.732	0.896
0.42	106	0.011	88.732	0.896
0.45	139	0.014	88.732	0.896

Table 2. The model performances at different learning rates.