Analysis of Logistic Regression and Linear Discriminant Analysis for Wine Quality Prediction and Breast Cancer Diagnostic

Adam Babs
School of Computer Science
McGill University
Montreal, QC H3A 0G4

Jizhou Wang
School of Computer Science
McGill University
Montreal, QC H3A 0G4

Nahiyan Malik
School of Computer Science
McGill University
Montreal, QC H3A 0G4

Abstract

The main objective of this project is to apply linear classification techniques to classify the quality of wine and the malignancy of breast cancer based on benchmark data sets. We applied logistic regression and Linear Discriminant Analysis (LDA) models on the data sets. Various methods have been applied to develop the algorithms and improve their accuracy. We demonstrate how data normalization, applying Chi-squared test and training the models using features and their interactions change the performance and accuracy. We found that the prediction accuracy does not differ much with logistic regression having a slightly improved prediction on both data sets (<1%). Although runtime wise, it is more than 10 times faster to train on LDA compared to logistic regression.

1 Introduction

The goal of this project is to classify wine quality and breast cancer malignancy. Both are classification problems; the wine quality is to be classified between good and bad and the breast cancer malignancy is to be classified between benign and malignant. Previous research based on these data sets (Lemionet 2015; Entezari 2013) state that classification using various methods works well on these data sets, with the breast cancer data set achieving much higher accuracy on testing sets, one of the main reasons being the high variance within the features. Comparatively, the wine data set exhibited much less variance on features.

Logistic regression and LDA were used to perform the classifications. The learning rate and a number of iterations were the main hyperparameters that were tuned for the logistic regression. LDA as a generative model had no hyperparameters for tuning. Its decision boundary for classification is based on the log-odds ratio of a gaussian distribution. The hyperparameters were directly calculated from the data set.

Feature normalization provided significant improvements in performance as well as accuracy in logistic regression. It had no substantial effect on the performance of LDA as it is a generative model discussed above. Additionally, interaction terms added based on chi-squared analysis were used with the wine data set to achieve a marginal increase in accuracy on logistic regression.

2 Data Sets

There were 699 entries provided in the breast cancer data set and of which 458 were benign tumors and 241 were malignant. We have removed 16 data entries as they were missing values on certain entries. This leaves us with 683 entries.

feature	mean	std	min	25%	50%	75%	max
clump_thickness	4.417740	2.815741	1.0	2.0	4.0	6.0	10.0
uniformity_of_cell_size	3.134478	3.051459	1.0	1.0	1.0	5.0	10.0
uniformity_of_cell_shape	3.207439	2.971913	1.0	1.0	1.0	5.0	10.0
margincal_adhesion	2.806867	2.855379	1.0	1.0	1.0	4.0	10.0
single_epithelial_cell_size	3.216023	2.214300	1.0	2.0	2.0	4.0	10.0
bare_nuclei	3.544656	3.601852	1.0	1.0	1.0	5.0	10.0
bland_chromatin	3.437768	2.438364	1.0	2.0	3.0	5.0	10.0
normal_nucleoli	2.866953	3.053634	1.0	1.0	1.0	4.0	10.0
mitoses	1.589413	1.715078	1.0	1.0	1.0	1.0	10.0

Table 1. Breast Cancer Data Set

In the wine quality data set, 1599 entries were split into a quality category scoring from 0 to 10. As we conducted binary categorization, the data entries were processed such that a quality score of 6,7,8,9,10 is categorized as positive and the rest as negative.

feature	count	mean	std	min	25%	50%	75%	max
fixed acidity	1599.0	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000
volatile acidity	1599.0	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000
citric acid	1599.0	0.270976	0.194801	0.00000	0.0900	0.26000	0.420000	1.00000
residual sugar	1599.0	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000
chlorides	1599.0	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100
free sulfur dioxide	1599.0	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000
total sulfur dioxide	1599.0	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000
density	1599.0	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369
pH	1599.0	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000
sulphates	1599.0	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000
alcohol	1599.0	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000

Table 2. Wine Quality Data Set

Initial examination of the data showed that the breast cancer data had higher standard deviations across features, which could lead to higher accuracy.

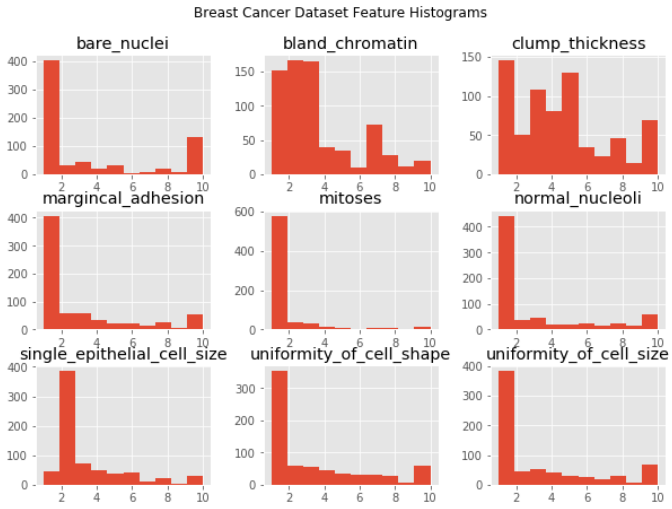

Figure 1. Breast Cancer Feature Histograms

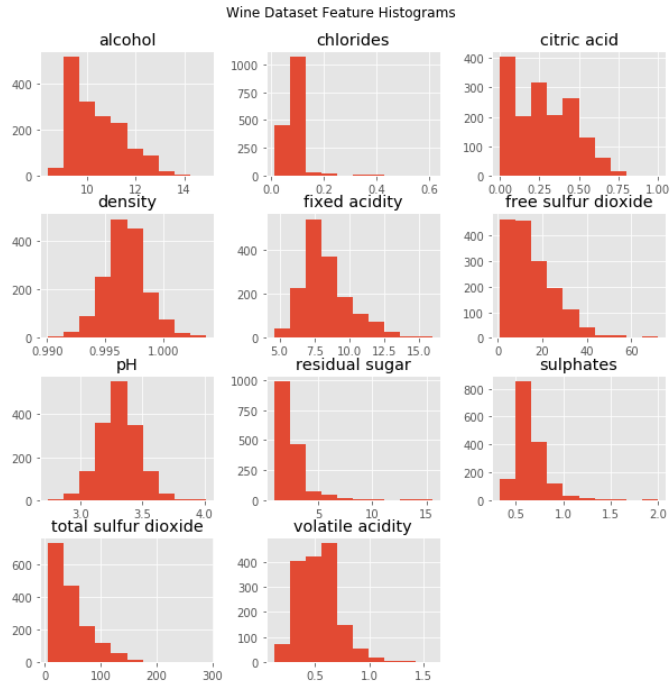

Figure 2. Wine Feature Histograms

The distributions of the breast cancer data set features are also much more skewed compared to the the wine data set features, which also explains the standard deviation disparities on Tables 1 and 2. The data sets were divided into training and validation sets and evenly split for 5-fold cross validation during the training-testing phase. The wine data set has more varied, continuous values, whereas the breast cancer data set is represented by integers in range 1-10. For feature selection in the wine quality data set, we tried training the logistic regression model using only the features which had positive correlation values against the labels, although this did not significantly improve accuracy of the trained model. Subsequently, we tried feature selection using chi- square test and adding interaction terms between features. These features included 2-way interactions terms up to 6-way interactions. Based on the Chi- squared test, we selected features based on a range of p-value cut-off shown in detail in the chi-squared test table below. Lower p-value cut-off shown in detail on Table 4. Lower p-value on feature suggests a higher dependence on class.

Ethical Considerations

In the botany world, 80% accuracy on iris classification might be sufficient, but to classify a mushroom as poisonous or edible for food, perhaps 99% (or higher) accuracy is required (Kiri L. Wagstaf, 2012).
Equivalently, while predicting a matter of a significant meaning such as breast cancer malignancy, at all costs, we should avoid misclassification as it may be detrimental. Ideally, we should aim for high sensitivity of the algorithm, whereas high accuracy when predicting the wine quality labels might be relatively less important.

Machine-learning models can learn the patterns of health trajectories of vast numbers of patients. This facility can help physician to anticipate future events at an expert level, drawing from information well beyond the individual experience and possibilities. However, The Institute of Medicine concluded that a diagnostic error will occur in the care of nearly every patient in his or her lifetime, and receiving the right diagnosis is critical to receiving appropriate care. (Alvin Rajkomar, Jeffrey Dean, Isaac Kohane 2019)

Therefore, we must be vigilant as to the extent we rely on algorithms knowing that misclassifications may have fatal consequences.

3 Results

There were several experiments that were run on both data sets using logistic regression and LDA. All accuracy metrics were measured using k-fold cross validation.

Logistic Regression Performance

There were two hyperparameters that were tuned for logistic regression: learning rate and the number of iterations. In order to first identify a suitable number of iterations,

a range of iterations were tried against different learning rates for the wine data set.

Learning rate	Iterations	Accuracy
0.1	1000	56.78389%
	10 000	59.40987%
	100 000	63.91536%
0.05	1000	56.84933%
	10 000	58.34894,%
	100 000	62.41261%
0.01	1000	56.90928%
	10 000	58.15948%
	100 000	62.102664%
0.005	1000	56.97237%
	10 000	58.34815%
	100 000	62.040164%
0.001	1000	57.47296%
	10 000	60.85011%
	100 000	64.03781%
0.0005	1000	56.97570%
	10 000	60.036442%
	100 000	62.91242%

Table 3. Learning Rates and Iterations

Although it appeared that 100000 iterations resulted in better accuracy, it did not guarantee the most efficient convergence speed based on the cross- entropy loss. In order to investigate if convergence could happen faster within 1000 iterations, we analyzed a range of learning rates on the data sets with min-max normalization.

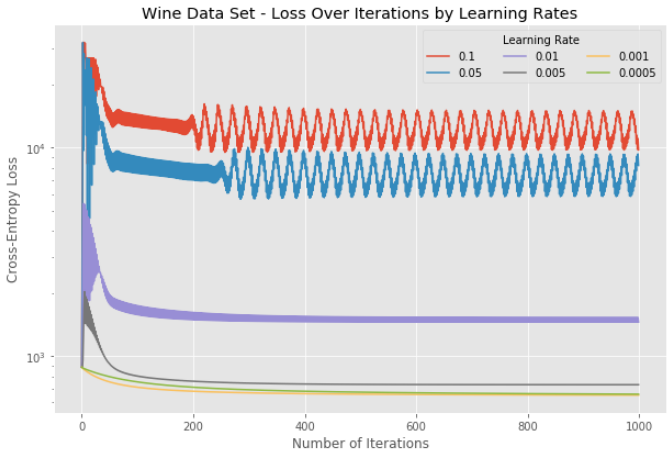

Figure 3. Loss Over Iterations by Learning Rates

It was found that convergence did, in fact, occur before 1000 iterations using normalized features. Larger learning rates such as 0.1, 0.05 and 0.01 resulted in weight oscillations. A learning rate of 0.001 exhibited the best convergence for both the wine and breast cancer data sets. As a result, for the rest of the experiments, a learning rate of 0.001 and 1000 iterations were used. The strategy used here as a stopping criterion for the number of gradient descent iterations was to look at the cross entropy loss over time and to observe an asymptotic decrease of the loss over iterations or until the total number of iterations was reached (IBM 2013). If the rolling slope (for example 100 iterations) of the loss function is flat, gradient descent could automatically be stopped.

Both the breast cancer and wine data sets benefited greatly from data normalization. Two normalization techniques were attempted: min-max and z-score.

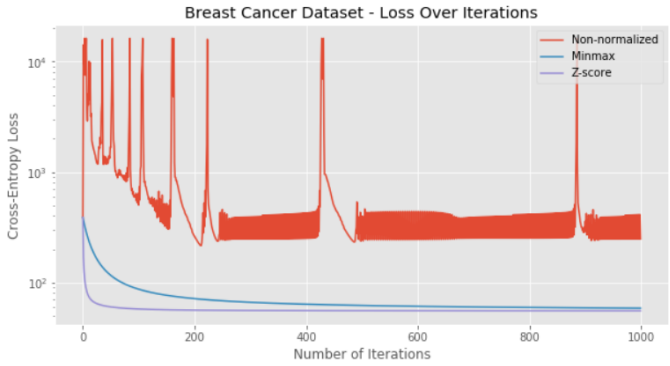

Figure 4. Breast Cancer Data Set - Loss Over Iterations

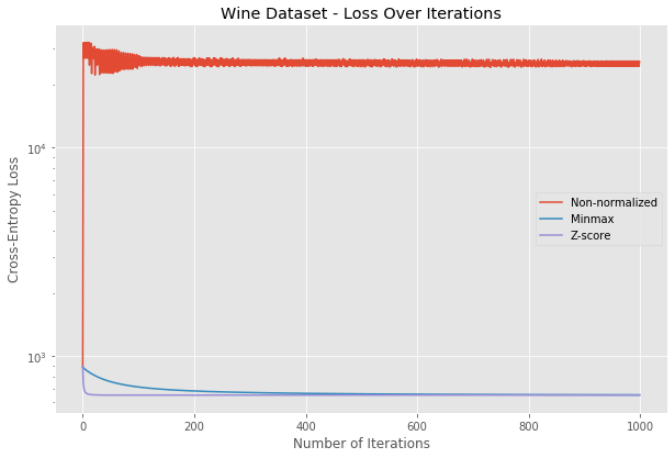

Figure 5. Wine Data Set - Loss Over Iterations

As can be observed from Figures 4 and 5, convergence with both normalized features vastly outperformed the unnormalized features. The best accuracy resulted from the min-max normalized features for both the breast cancer and wine data data sets, with 96.55% and 74.31% respectively.

Feature interaction experiments for logistic regression on Wine Quality data set

Bias is an essential aspect of model selection. (Mitchell, 1980) As our model becomes more complex with an increasing number of feature/interaction, its variance also increases. The trade-off between simple and complex model is akin to a trade-off between a model with high bias/low variance compared to a model with low bias/high variance. We want to find an optimal model with the least validation error.

To minimize this error, chi-squared tests were conducted on feature selection. From Table 4 shown below, by default having 11 features gave us an accuracy of 0.7431.

In simple models with interactions up to 1,2,3-way, the lowest feature count had the lowest accuracy shown in red. This suggested underfitting. In contrast, the highest feature count (including all the interaction terms) gave the highest accuracy.

In complex models with interactions up to 4,5,6-way, the highest feature count had the lowest accuracy shown in red. This suggested overfitting. Intriguingly, the highest accuracies selected were of models with around 300-400 features with the highest being 0.7536. This is an increase in accuracy of about 1% compared to having 11 features by default.

Chi-squared p-value Features Selection						
p value	1-way	features	2-way	features	3-way	features
0.01	0.738768	3	0.742206	24	0.74253	56
0.05	0.742025	5	0.743534	33	0.74427	83
0.1	0.742025	5	0.744331	38	0.74641	103
0.25	0.742464	6	0.742974	43	0.74378	135
0.5	0.742149	9	0.743087	51	0.74365	166
1	0.743087	11	0.744471	66	0.7474	231
p value	4-way	features	5-way	features	6-way	features
0.01	0.745034	105	0.743905	152	0.74477	180
0.05	0.748711	161	0.75035	254	0.75071	337
0.1	0.748898	209	0.753905	338	0.7536	457
0.25	0.75041	303	0.742779	518	0.73659	719
0.5	0.750093	391	0.719079	704	0.70745	1024
1	0.718076	561	0.689679	1023	0.68462	1485

Table 4. Chi-squared p-value Feature Selection

Runtime and Accuracy Comparisons

Runtime and accuracy tests were conducted on both data sets with normalized base features. The hyperparameters for logistic regression were set at a learning rate of 0.001 and for 1000 iterations.

The average execution time for logistic regression for the wine data set was 0.1244 (sd 0.0235) seconds. The average runtime for the breast cancer data set was 0.0776 (sd 0.0169) seconds. The decrease in average time for the breast cancer data set corresponds to the decrease in the size of the data set.

LDA had a substantial increase in runtime on both data sets. The average execution time on the wine data set was 0.00691 (sd 0.00150) seconds. The average runtime for the breast cancer data set was 0.00300 (sd 0.000945) seconds.

Runtime

Model	Wine	Cancer
LDA	0.00691 (sd 0.00150) seconds	0.00300 (sd 0.000945) seconds
Logistic Regression	0.1244 (sd 0.0235) seconds	0.0776 (sd 0.0169) seconds

Table 5. Runtime comparisons of the implemented algorithms

For accuracy, both models provided similar results in 5-fold cross validation shown in the table below.

Accuracy

Model	Wine	Cancer
LDA	74.27% (sd 0.0225)	96.04% (sd 0.0150)
Logistic Regression	74.30% (sd 0.0223)	96.55% (sd 0.01140)

Table 6. Accuracy comparisons of the implemented algorithms

4 Discussion and Conclusion

Logistic regression and LDA both had similar results in terms of accuracy, which was expected as they are both linear classification models. For both models, normalizing breast cancer and wine data sets provided the most drastic improvements in accuracy. For the wine data set, adding interaction terms did slightly improve the accuracy, but there is a risk of overfitting as more interaction features are added to the model.

For both data sets, principal component analysis (PCA) could be applied to the features to reduce the number of dimensions (Suleiman 2014). This could be especially useful to apply on the additional interaction terms substituting the chi-square test as the number of interaction terms increase. The logistic regression model could also be improved with decaying learning rates. As discussed above, larger learning rates have a tendency to oscillate weights, but by measuring the rolling slope of the loss function, the learning rate could be decreased over iterations, starting from a larger number. Similarly, the gradient descent could be terminated earlier if the slope of the loss function is under a certain threshold. If the wine and breast cancer data sets were larger, stochastic gradient descent could also be implemented for faster convergence. These implementations could potentially increase both performance and accuracy.

As for LDA, we have assumed that the covariance matrix is the same for all classes. It may be more optimal to assume the covariance matrix is distinct between classes. For future investigation, it is possible to test out different implementations of the Bayes rule model such as QDA and Gaussian Naive Bayes if features are strongly independent.

References

[1] Mitchell, Tom M. 1980. The Need for Biases in Learning Generalizations. CiteSeerX.

[2] Wagstaff, Kiri L. 2012. Machine Learning That Matters.

[3] Lemionet, Amelia; Liu, Yi; Zhou, Zhenxiang. 2015. Predicting quality of wine based on chemical attributes. Stanford University.

[4] Entezari, Reihaneh. 2013. Breast Cancer Diagnosis via Classification Algorithms. University of Toronto.

[5] IBM. 2013. Stopping Criteria (complex samples logistic regression algorithms). IBM.

[6] S. Suleiman; Issa, Suleman; U. Usman; Salami,Y. 2014. Predicting an Applicant Status Using Principal Component, Discriminant and Logistic Regression Analysis.

[7] Alvin Rajkomar, Jeffrey Dean, Isaac Kohane. 2019. Machine Learning in Medicine.