Analysis of Logistic Regression and Linear Discriminant Analysis with Implementations on Real-World Datasets

Christopher Zheng[1], Lihui Huang[2] and Haoxuan Shi[3]

***Abstract*— Linear classifiers have been extensively explored. In this work, we focus on Logistic Regression (LR) and Linear Discriminant Analysis (LDA) by applying them to two public datasets and comparing their accuracy and efficiency scores of undertaking binary classification tasks. We study how the LR parameters can impact the prediction accuracy and the model's convergence rate which is strongly associated with the overall runtime. We also present how certain data processing and feature selection, achieved by cross-correlation analysis, can contribute to the accuracy boost and runtime reduction. Specifically, our correlation analysis selects four features out of eleven from one of the datasets and we generate a new feature based on some domain knowledge. This feature combination, along with proper feature normalization, enhances two models' prediction accuracies to 75.1% (LR) and 96.3% (LDA), respectively, based on K-fold cross validation. As a result, our procedures can boost the prediction accuracy by up to 19% on selected datasets.**

I. INTRODUCTION

Background

Linear classification, one of the most salient applications of modern day machine learning, has been extensively studied. We implement two linear classification algorithms (Logistic Regression [1] and Linear Discriminant Analysis [2]) and let them classify example entries of two datasets.

The first dataset is the red wine quality dataset ("wine dataset" in short) [3] which provides 11 features along with a grade of quality for each example. The features are all characteristics associated with the wines, e.g., the concentration of citric acid and of sulphates.

The second dataset is the Breast Cancer Wisconsin (Diagnostic) Dataset ("tumor dataset" in short) [4]. The tumor dataset includes 9 features as well as ground truth labels. These features physically describe the tumors in terms of, for instance, their sizes and shapes.

Project Goal

As the purpose of this work, we plan to implement and use the two aforementioned linear classifiers to analyze and learn the features of the wines and tumors, and then predict whether some wines are considered good or not and whether some tumors are benign or not.

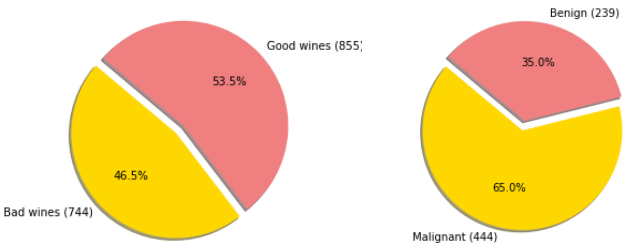
Generally, for n input examples, suppose we have m features $\{x_1, x_2, ..., x_{n-1}, x_m\}$ and one binary class c for each one of them. The task of classifiers is to produce a set, with length m, of weights $\{w_1, w_2, ..., w_{n-1}, w_m\}$ and the prediction model takes shape: $y = w_1x_1 + w_2x_2 + ... + w_{n-1}x_{n-1} + w_nx_n$, where y is the predicted class.

Related Work

Two general categories of models are commonly used to determine the expressions for linear classifiers: *generative models* and *discriminative models*. Well-known generative models include Linear Discriminant Analysis (LDA) and Naive Bayes classifier (NB) [5]. LDA, commonly utilized for dimensionality reduction before later classification, finds a linear combination of features to separate classes. NB is a probabilistic model that uses Bayes' Theorem assuming strong conditional independence. NB converges more quickly than discriminative models (e.g., Logistic Regression) do under that assumption.

On the contrary, discriminative models, such as Logistic Regression (LR), Perception [6] and Support Vector Machine (SVM) [7], are also widely used. Logistic Regression estimates parameters of a logit model and classifies examples

[1]Christopher Zheng (260760794, Corresponding Author): christopher.zheng@mail.mcgill.ca
[2]Lihui Huang (260821232): lihui.huang@mail.mcgill.ca
[3]Haoxuan Shi (260779480): haoxuan.shi@mail.mcgill.ca



(a) Wine dataset (b) Tumor dataset

Fig. 1: Proportion of classes in two datasets.

between classes based on probability. Perception is a binary classifier that may potentially fix some problems occurred in the training set. SVM is an algorithm that tries to maximize the distance or margin between the entries in the training set and the determined hyper-plane.

Previous works [3, 8] on the same wine quality dataset explore various machine learning algorithms. Besides, detailed feature selection techniques in the pre-processing stage using statistical knowledge like correlation analysis have been studied in [9].

II. Datasets

We briefly clean both the wine dataset and tumor dataset by removing examples that contain non-numeric or null feature values. To keep datasets inclusive enough, we do not remove outliers from examples. The proportions of positive and negative classes are illustrated in Fig. 1. It is clear that neither one of the two datasets is unacceptably imbalanced, therefore, we proceed to the feature selection step.

Wine Dataset

The wine dataset contains 1599 valid red wine examples, 11 features and a quality class for each. To select the most representative features, we undertake a correlation analysis and the result is shown in Fig. 2.

1) *fixed acidity*: not selected. This feature has a correlation of 0.67 with citric acid that has higher correlation with the quality than fixed acidity does.
2) ***volatile acidity***: selected. Although it is -0.55 correlated with citric acid, it has strong correlation with the quality.

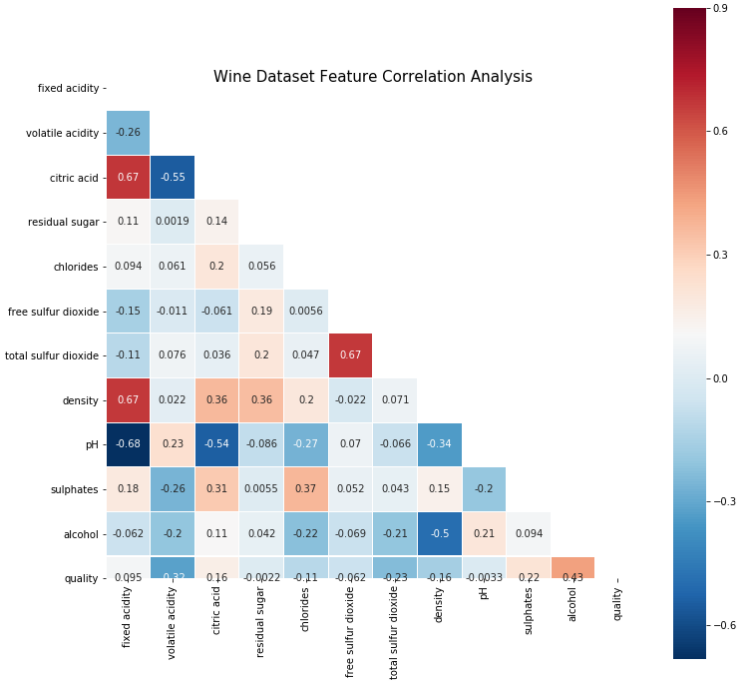


Fig. 2: Correlation analysis of features of the wine dataset. This contributes to the feature selection of the wines. (Visualization inspired by [10].)

3) ***citric acid***: selected as explain in 1.
4) *residual sugar*: not selected due to overly weak correlation with the quality.
5) chlorides: not selected as in 4.
6) *free sulfur dioxide*: not selected for the same reason as 4 even though it has quite strong correlation with the total sulfur dioxide as the latter breaks down to the former.
7) *total sulfur dioxide*: not selected for the same reason as 6.
8) *density*: not selected since it has the cross correlation with alcohol which is more important to wines and thus selected instead.
9) *pH*: not selected due to the cross correlation with citric acid which is selected in 3.
10) ***sulphates***: selected because of its high correlation score with the quality.
11) ***alcohol***: selected as discussed in 8.

In sum, the most representative features are the volatile acidity, citric acid, sulphates and alcohol. Besides, we generate a new feature based on some basic chemistry knowledge [11]. If there exists not enough SO_2 in wine, the quality score will

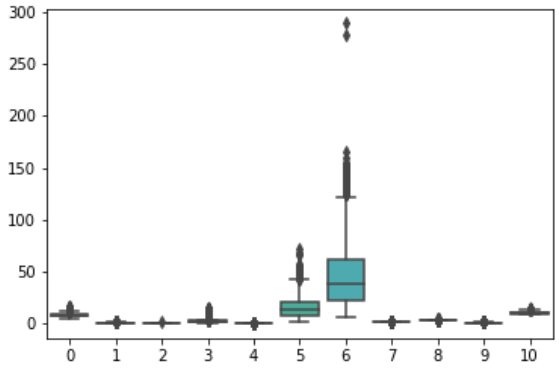


Fig. 3: Distribution summary of the features values of the wine dataset.

gradually decrease and higher pH does require more free sulfur dioxide. Using molecular SO_2's formula,

$$MSO_2 = \frac{free\,sulphur\,dioxide}{1 + 10^{pH-1.81}}$$

The extent to which this feature selection improves our models is discussed in Section III.

As the last step, we normalize the features so that all of them fall into the range of zero to one. We include this step of normalization because, for example, the distributions of the features of the wine dataset vary to a quite large extent as in Fig. 3.

Formally, let n denote the size of the dataset, $\forall f_i \in F$ the whole feature space, $\forall$ value $x_j \in f_i$,

$$x_j = \frac{x_j - min(x_1, x_2, ..., x_n)}{max(x_1, x_2, ..., x_n) - min(x_1, x_2, ..., x_n)}$$

By doing so, we can initially balance the influence of each feature so that the models (especially the Logistic Regression) can produce higher accuracies.

Tumor Dataset

The tumor dataset contains 683 usable examples of tumors, 9 features and one class label for each. We again analyze the cross-correlations among features and the result is shown in Fig. 4.

Similar to what we do in the wine dataset, the feature selection of the tumor dataset excludes three insignificant features (cell size, cell shape, number of bare nuclei) and keeps six suggestive

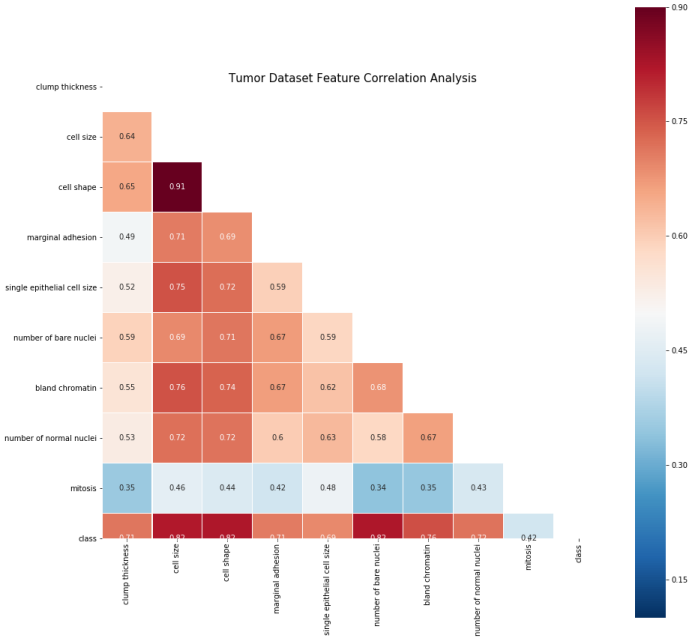


Fig. 4: Correlation analysis of features of the tumor dataset. This contributes to the feature selection of the tumors. (Visualization inspired by [10].)

features. Finally, we complete the normalization step and finish the pre-processing stage.

Ethical Concerns

As machine learning practitioners, we may gain access to or infer some confidential or private information which we should be working on carefully and ethically [12]. Precisely, for the wine dataset, the detailed information of the wines and their corresponding classes may consist of certain trade secrets of wine companies. Disclosure of such information for non-academic purposes may inevitably impact their revenues. For the tumor dataset, the feature values (e.g. IDs) may be linked to specific patients and we must respect their privacy including gender, sex, age, occupation, etc. Briefly, it is of essence to protect our research data.

III. Results

We commence by comparing the runtime performance of LR based on different learning rates. Fig. 6 suggests that large learning rates can result in forever divergence and low accuracy, while learning rates that are too small can take a long time for the model to converge. This issue causes inefficiency. In our case, we decide to use the learning rate of 0.01 for all of our following

ACCURACY	LR w/o normalization	LDA w/o normalization	LR w/ all features	LDA w/ all features	LR w/ 4 feat. selected + 1 new	LDA w/ 4 feat. selected + 1 new
Wine dataset	0.565	0.724	0.742	0.738	**0.751**	**0.749**
Tumor dataset	0.943	0.951	0.931	0.95	**0.963**	**0.968**

TABLE I: Accuracy comparison of classifications with and without input normalization, and of basic classifications with all available features as input and improved classifications with selected features plus one newly generated feature.

experiments since it has the advantage of fast converging and achieving high accuracy.

Next, we compare the efficiency of LR and LDA. According to Table I, the accuracy of LR on the wine dataset is greatly improved from 56.5% to 74.2% after all the features are normalized. Therefore, the normalization can reduce the impact of the differences among ranges of distinct features and thus improve accuracy. After we normalize the features and set an appropriate iteration number for the LR model, both models achieve similar accuracies on two datasets. A significant discrepancy, however, exists between the runtime of two models. Considering the wine dataset, Fig. 5 indicates that the LR model with an iteration number of 40 takes 145 milliseconds to converge while the LDA model takes only 6 milliseconds to finish. This implies that LDA is a much more time efficient model than LR.

Finally, we enhance our prediction accuracy by selecting a subset of features and adding a new feature to train models for both datasets. We select the representative features and remove the features that are correlated and have relatively less contribution to the quality of wine. Based on the correlation analysis (as discussed in Section II), when we encounter two features that have a high correlation, we keep the feature that the quality is more dependent on and remove the other. For the wine dataset, we select four features, generate a new feature, and form our set of features for training. According to Table I, our model accuracies are to some extent improved on both datasets because of this feature selection.

Note: all the prediction accuracies are generated using K-fold cross validation to ensure that the evaluation result are relatively accurate.

[1]Implementation specs: models of both Logistic Regression and LDA are implemented in Python 3.7.3 and our studies are conducted on a PC with Intel i7 CPU and a 8 GB RAM.

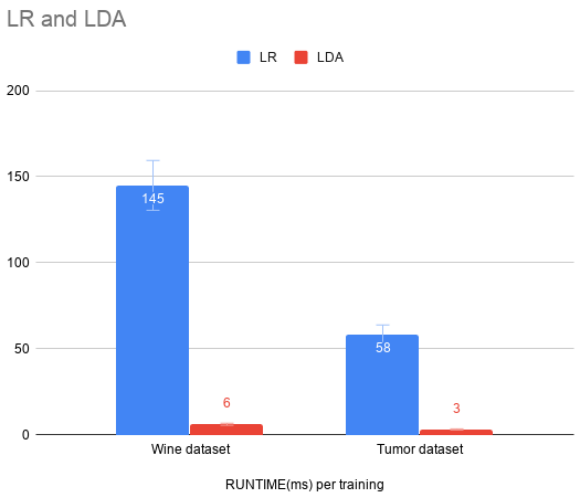


Fig. 5: Runtime [1]comparison of Logistic Regression (LR) and Linear Discriminant Analysis (LDA) on both the wine dataset and tumor dataset.

IV. Discussion and Conclusion

To improve the runtime efficiency of Logistic Regression, we design a stopping criteria for LR models. First, we fix our learning rate and estimate a maximum number of iterations that allow the model to converge and meet an acceptably high accuracy. Second, we define a threshold to determine a maximum difference between the accuracies of different iteration numbers. If the accuracy improvement is above the preset threshold, we proceed to the next iteration. When the accuracy improvement is below our defined threshold, we stop our training process. This stopping criteria prevents the LR model from continue to run while making ignorably little progress on improving the accuracy. The validation of this stopping criteria is left for future work.

In this work, our two models are implemented to solve binary classification problems. Nonetheless, in real-word scenarios, we may inevitably encounter situations where more than 2 classes are present. This leads us to the use of Quadratic

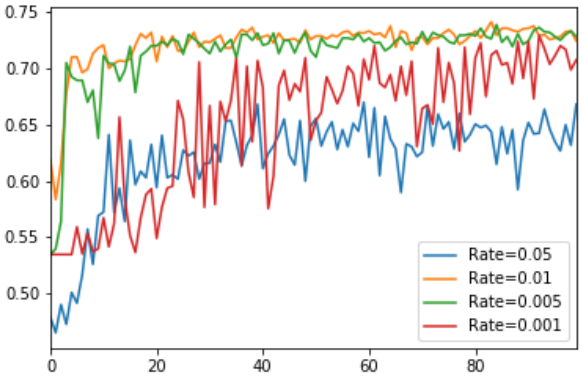


Fig. 6: Accuracy and trend comparison (y-axis) of Logistic Regression with different learning rates with respect to an increasing number of iterations from zero to one hundred (x-axis).

Discriminant Analysis model (QDA), which allows different co-variance matrices for each class. Moreover, we have to consider the situations where the features are no longer considered linear. In practice, we might add different exponents to our features and accordingly develop non-linear estimation functions.

In conclusion, we implement two linear classifiers, Logistic Regression and Linear Discriminant Analysis, and closely study their classification accuracies and runtime efficiencies on two publicly recognized real-world datasets. We research to what extent specific feature selection and engineering techniques and data processing strategies (e.g., normalization) can increase the overall classification accuracies. We also demonstrate that the hyper-parameters can noticeably impact the performance of the Logistic Regression model, and naturally we propose a straight-forward solution to make the model converge more quickly.

V. Statement of Contributions

1) Christoper Zheng. Implementation: responsible for pre-processing, feature selection, accuracy evaluation and k-fold cross-validation. Write-up: responsible for the Abstract, Introduction, Datasets, Statement of Contributions, graphs and tables.
2) Lihui Huang. Implementation: responsible for programming the LDA model. Write-up: responsible for the Results, and Discussion and Conclusion.
3) Haoxuan Shi. Implementation: responsible for programming the LR model and studying the convergence criteria of it.

References

[1] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1-2, pp. 167–179, 1967.
[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
[3] P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Using data mining for wine quality assessment," in *International Conference on Discovery Science*, pp. 66–79, Springer, 2009.
[4] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical image processing and biomedical visualization*, vol. 1905, pp. 861–870, International Society for Optics and Photonics, 1993.
[5] M. E. Maron, "Automatic indexing: an experimental inquiry," *Journal of the ACM (JACM)*, vol. 8, no. 3, pp. 404–417, 1961.
[6] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine learning*, vol. 37, no. 3, pp. 277–296, 1999.
[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[8] Y. Er and A. Atasoy, "The classification of white wine and red wine according to their physicochemical qualities," *International Journal of Intelligent Systems and Applications in Engineering*, pp. 23–26, 2016.
[9] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," *Procedia Computer Science*, vol. 125, pp. 305–312, 2018.
[10] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
[11] S. Rosa, "S02 calculation," Accessed: 2019-09-28.
[12] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.