COMP 551 Applied Machine Learning Miniproject 1 Report

Zhi Wen, Yunfei Cheng, Shih-Chieh Fuh

Abstract

In this project, we aim to retrieve data from the two datasets (wine quality and breast cancer diagnosis), apply logistic regression and linear discriminant analysis (LDA) on both datasets, and finally compare the two methods. Without particular data processing, the LDA produced 73.11% and 95.91% accuracy, whereas the logistic regression model produced 72.98% and 86.41% accuracy for wine and cancer data respectively. Furthermore, special attention was given to how the model learned in the logistic regression model. A thorough investigation to find the best combination of factors: number-of-iterations, initial-learning-rate, and learning-rate-decay was conducted. We found that on both datasets LDA performs better than logistic regression while being faster to run. We also analyzed features' importance in deciding models' performance and justified our analysis by using only the most important feature selected by our analysis.

Introduction

In this project, our tasks are to acquire, preprocess, and analyze the two data sets of wine quality and breast cancer diagnosis. We apply two techniques: Logistic regression and LDA, to build up classification models for these two datasets. On the wine quality dataset, we managed to achieve 72.98% cross validation accuracy with logistic regression and 73.11% with LDA. This is comparable with others' work on the same dataset using Feedforward NN and SVM.[1] On the breast cancer dataset, we managed to achieve 86.41% cross validation accuracy with logistic regression (76.32% precision and 83.72% recall) and 95.91% with LDA (96.76% precision and 92.35% recall), which is comparable with other work on the same dataset.[2] In addition, we included customizable decision boundary as a parameter of our models in order to accommodate the possibility in health-related tasks that recall and precision are of different importance.

Results above are achieved by exhaustively searching for optimal combination of hyperparameters for logistic regression, which is detailed below. In addition, an analysis on attributes' relative importance in deciding classification accuracy was done. Basically it tries to assign importance to attributes by examining the difference of their average across classes.

Dataset

The wine dataset is first separated into two groups, positive (quality score 6-10) and negative (quality score 0-5), which are then converted to a binary classification as 1 and 0 respectively. There are 855 entries in the positive group and 744 entries in the negative group. The attributes in different scale are converted to percentile score. From the mean of the positive/negative sets, we find that alcohol, sulphates, and volatile acidity are the top 3 attributes with the greatest difference in average percentile between the two groups. The positive group tends to have high alcohol, high sulphates, and low volatile acidity score.

The breast cancer diagnosis dataset contains columns of patient ID, 9 attributes, and a category (number 2 and number 4 for benign and malignant, respectively). There are 444 entries in the benign group and 239 entries in the malignant group, with 16 entries excluded because of the missing data. Since the dataset are in the same scale, we directly compare the average score between benign and malignant for

each attribute. For all attributes, the malignant group has higher scores. The top three attributes with greatest difference are uniformity of cell size, uniformity of shape, and bare nuclei. In addition, one thing to notice is that the range of average score from the benign group is between 1-3, and the range of average score from the malignant group is between 5-8, for all attributes except for mitoses. For mitoses, both groups have low average scores. The two groups may indeed have the similar low scores; however, it is possible that the score scale / criteria of mitoses are not well defined. For example, a scale may cover a range significantly broader than the range covered by the datasets. In such case, the data may fall into a limited range of the score, which may explain the similar low scores for mitoses.

By transforming the wine quality data into percentile ranking, we put all the attributes into the same scale. The attributes from the breast cancer data are all in the same scale, so we choose not to transform the data. By comparing the average score between the two groups, we have selected different subsets of features to compare with all features, and a selected few features may be enough for the classification tasks. Take the breast cancer dataset as an example: The accuracy by LDA is off by only 1% with only the top three features provided (data not shown). With a large amount of features, the time required for analysis may be too long. Although intuitively, the more features and information provided, the more accurate a model may be, the result from top features seems promising. With the preprocessing and categorizing the features into different ranks, in terms of the difference in average score between the two groups, perhaps selecting only the most influential features may yield a satisfactory result.

While studying machine learning, certain ethical concerns may arise with the nature of the datasets. Take the wine dataset as an example, when studying a dataset that may be related to goods and profit, it is crucial for the researchers related to declare potential conflicts of interest beforehand. With health care data such as the breast cancer dataset, the participants should be kept anonymous for privacy concerns.

Results

The two figures at the bottom of figure 1 show differences of features across classes, and such differences can serve as indicators for corresponding feature's importance in downstream classification task. The bottom right figure shows that feature 6, 2, 3, and 8 have greatest differences across classes in the given order, and experiments show that using only feature 6 achieves 89.61% accuracy with LDA on breast cancer data, compared to 95.91% with all features. The small margin between the numbers suggests that feature 6 is indeed playing an important role in predicting whether a tumor is benign or malignant. Moreover, since each feature is well defined in data's specification, analysis conducted in our work that examines features' differences across classes can be considered as a way of finding strong predictor of a certain task. In our case, feature 6 is bare nuclei according to data specification, and this finding is in accordance with other studies.[3]

Similar auxiliary investigation was conducted on wine quality data. Specifically, the bottom left figure indicates that feature 11 is of great importance in determining whether a wine is good or not. Using only feature 11 with LDA achieves 69.92% accuracy, compared to 73.11% when using all features. This further justifies the use of our analysis.

For logistic regression, we investigated in depth how different combinations of hyperparameters impact model's cross validation accuracy. For decay, we experiment with a range of 1e-9 to 0.1. For

learning rate, we test with a range of 1e-7 to 10. The accuracy peaks at decay equalling 1e-6 for the wine quality dataset, while the breast cancer dataset has an increasing trend with smaller decay. Both the wine quality dataset and the breast cancer dataset show an increasing trend with the increase of learning rate and peak at 0.01 and 0.001, respectively. The change of accuracy between different settings in the wine quality dataset is relatively small compared to the breast cancer dataset, and the increase is more significant in lower learning rates but gradually plateaued. And as expected, number of iterations is positively, though weakly, correlated with overall performance, as can be found in the following figures.

We also note that both learning rate and decay are crucial for achieving best performance. This is shown by the fact that for wine quality data the top 36 configurations all have decay of 10^{-7} with varying learning rate and number of iterations, and that for breast cancer data 11 out of the top 13 configurations have learning rate of 10 and 10 out of the top 13 have decay of 10^{-5} .

Unsurprisingly, LDA is much faster to run compared to logistic regression, since it has no need for iteratively updating parameters. We compared the runtime of LDA on breast cancer data and runtime of logistic regression with learning rate of 10 and decay of 10^{-5} , trained with 2000 iterations. On a 2019 MacBook Pro with 1.4 GHz Intel Core i5, it took LDA 1.848s to run while it took logistic regression 1.061s to run.

Discussion and Conclusion

This project aims at implementing logistic regression and linear discriminant analysis for classification tasks. We tested the two models on two separate datasets and further investigated a range of hyperparameters' impact on models performance. Our work shows that LDA outperforms logistic regression on both datasets while taking less time to run. However, as reviewed in the work of Anagnostopoulos et. al., the performance of LDA on the breast cancer diagnosis dataset is not among the best methods.[4] Our future direction may include investigating the top ranking methods and compare their strengths and weaknesses.

Statement of contributions

Zhi Wen implemented LDA model and wrote scripts for cross validation and selecting optimal hyperparameters. He also investigated models' performance when using fewer features and helped implementing logistic regression.

Yunfei Cheng implemented the logistic regression model with Zhi Wen's help and implemented script to read and preprocess wine data.

Shih-Chieh Fuh was responsible for the data logistics and graphical analysis, and completed most of the write-up along with Zhi Wen.

Reference

[1] Cortez, Paulo, et al. "Modeling wine preferences by data mining from physicochemical properties." Decision Support Systems 47.4 (2009): 547-553.

[2] Šter, Branko, and Andrej Dobnikar. "Neural networks in medical diagnosis: Comparison with other methods." International conference on engineering applications of neural networks. 1996.

[3] Narasimha A, Vasavi B, Kumar HM. Significance of nuclear morphometry in benign and malignant breast aspirates. Int J Appl Basic Med Res. 2013;3(1):22–26. doi:10.4103/2229-516X.112237

[4] Anagnostopoulos, Ioannis, et al. "The Wisconsin breast cancer problem: Diagnosis and TTR/DFS time prognosis using probabilistic and generalised regression information classifiers." Oncology reports 15.4 (2006): 975-981.

Figures

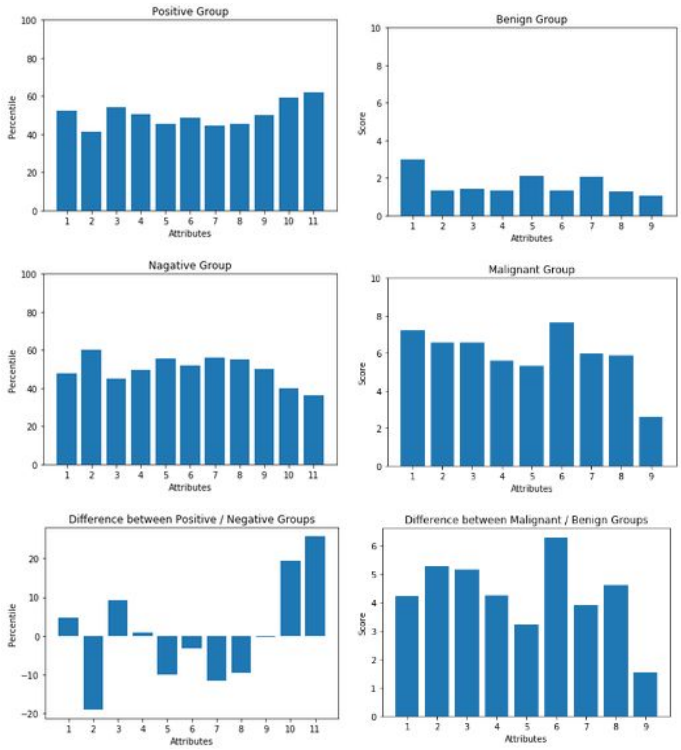

Figure 1. Average score comparison of wine quality (left) and breast cancer diagnosis(left). The traits are numbered in the X-axis. The traits 1-11 in the wine quality dataset are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The traits 1-9 in the breast cancer diagnosis dataset are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. In the wine quality column, the Y-axis is percentile. The top and the middle are the average percentile of the positive / negative group, respectively. The middle is the average percentile of the negative group. The bottom is the difference between (positive percentile – negative percentile). In the breast cancer diagnosis column, the Y axis is score (1-10). The top and the middle are the average score of the benign / malignant group, respectively. The bottom is the difference between (benign percentile – malignant percentile).

Decay	1.00E-09	1.00E-08	1.00E-07	1.00E-06	1.00E-05	1.00E-04	1.00E-03	1.00E-02	1.00E-01
WQ	0.587557	0.593046	0.654806	0.671857	0.634969	0.607658	0.598548	0.595284	0.594937
BC	0.791015	0.793594	0.767522	0.778878	0.758008	0.73443	0.702973	0.674472	0.651567

LR	1.00E-07	1.00E-06	1.00E-05	1.00E-04	1.00E-03	1.00E-02	1.00E-01	1.00E+00	1.00E+01
WQ	0.560929	0.607768	0.622009	0.616393	0.616094	0.630949	0.628187	0.626789	0.629544
BC	0.422271	0.58261	0.69604	0.819891	0.846038	0.828252	0.817162	0.817085	0.823109

Step	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000
WQ	0.594508	0.604248	0.608568	0.61103	0.617323	0.622878	0.623273	0.624656	0.623793	0.623792
BC	0.722084	0.731872	0.738109	0.739796	0.741952	0.744869	0.740906	0.741677	0.745103	0.745252

Figure 2. Accuracy with different values of decay (top), LR(learning rate, middle), and step in logistic regression. WQ: Wine quality dataset. BC: Breast cancer diagnosis dataset.

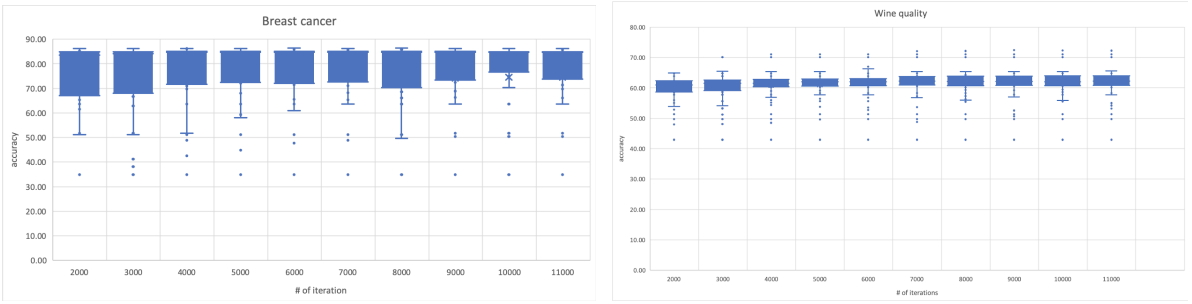

Figure 3. Boxplots of cross validation accuracy versus number of iterations, for breast cancer dataset and wine quality dataset.