COMP 551 Project 1
Machine Learning 101

September 28, 2019

(Group 65)
Hongshuo Zhou 260792817
Weige Qian 260763075
Yichao Yang 260764629

Abstract

In this project, we investigated the performance of two versions of linear classification models on two benchmark datasets, with logistic regression approach and LDA(linear discriminant analysis) approach. We examined the given features of both data sets and also new features of the Wine quality dataset. In particular,"**fourth root**" considerably improved the performance of the model. For model training and validation part, we found that on Breast Cancer Wisconsin (Diagnostic) Data Set, the LDA approach was achieved slightly lower accuracy than the logistic regression approach. While on Wine Quality Data Set, the LDA approach and the logistic regression approach was achieved the same level accuracy. For both two data sets, the logistic regression approach was greatly slower to train.

1 Introduction

1.1 Dataset information

Human wine taste could be predicted based on easily available analytical tests at the certification step; the dataset we used is related to red wine of the Portuguese"Vinho Verde" wine. For breast cancer dataset, features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image.

1.2 Task Description

Generally, we implemented both the logistic regression approach and LDA(linear discriminant analysis)approach on the datasets and made comparisons between the two models from accuracy(prediction error) and running time.

Previous work in wine quality prediction mainly focused on closely related and directly available analytical features from tests. In contrast, there exist other factors that also generate influence on the quality of wines. In this project, we inspect the effect of multiple features such a"fixed acidity" "free sulfur dioxide" and "residual sugar" for wine quality dataset, "radius (mean of distances from center to points on the perimeter) ","texture (standard deviation of gray-scale values) " and "perimeter" for breast cancer dataset. Additional features for wine quality dataset"**fourth root**" are introduced.

1.3 Important findings

From the project, we had two important findings. First, we assumed that logistic regression approach should be a better model for both of the datasets, as it includes iterative update operations. However, based on our experiments, its actual performance on both datasets was not as good as our expectations. Precisely saying, it took much more time than the LDA approach to get exactly the same accuracy on the wine quality dataset, and slightly fell behind LDA approach on the breast cancer dataset. But this should not happen when we have much larger datasets, as LDA involves computing transpose and inverse matrix. If researchers considering accuracy more important than runtime on larger datasets, it is recommended to use logistic regression approach as it should behaves more stable and accurate. However, researchers still need to choose between those two models because LDA approach has a unique advantage on runtime.

1.4 Previous works

Past work on wine quality [1] already proposes a data mining approach to predict human wine taste preferences, using the support vector machine approach, the multiple regression approach and the neural network approach.

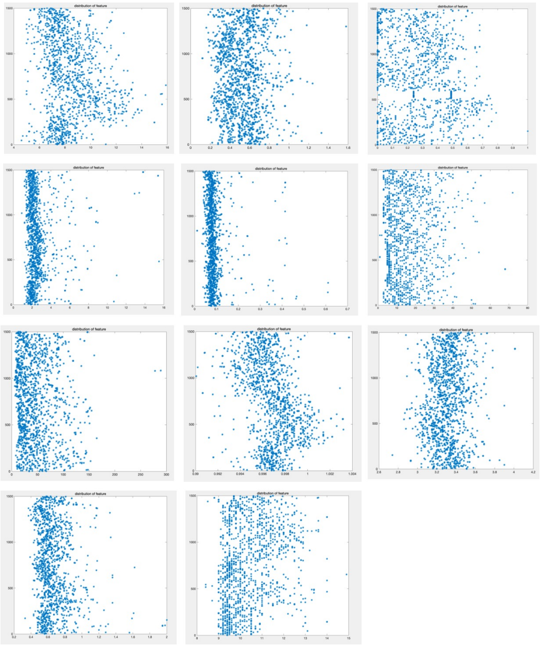
2 Dataset

2.1 Size information

The total size of the wine quality dataset is 1600, and it is a data sheet containing the related features of red wine. The total size of the breast cancer dataset is 699; it is a data sheet containing the related features of each cell nuclei from the fine needle aspirate (FNA) of breast mass. Instead of splitting the whole dataset by the ratio 10:1:1 to form a training set, validation set and test set, we choose k-fold cross validation method to split the original dataset into training/validation set for each model.

2.2 Given features and processing

For our desired features matrix, we first removed all the samples in breast cancer datasets that with missing/malformed features. Then we transformed the "quality" feature into binary form by split the quality ratings beyond 5 and less than or equal to 5, and all other ratings treated as negative. Finally, we concatenated feature data arrays from two datasets into two corresponding matrices use for training and prediction. We used a different method to analyze our datasets. For example, by plotting the dependence between two features, we can easily find out if they are linearly dependent. Those plots below are the visualization of the distribution of all input features. We can see clearly that most features have dispersed distributions while some features have closed value among observations. This provides a breakthrough for finding new features



(a)

Figure 1: Distributions plotted in all 11 wine quality features.

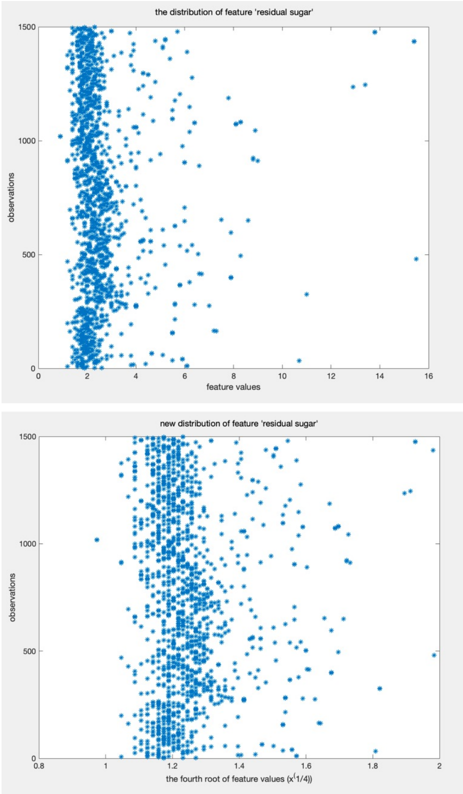
2.3 New features

2.3.1 Subset of features

We generated a new subset of feature by deleting features“residual sugar”, “chlorides”, “free sulphur dioxide” and “total sulphur dioxide”.

2.3.2 Additional features

We created a new feature called **“fourth root”** by taking the fourth root of the original feature “residual sugar,” the figure is the distribution of that new feature compare to the original feature. The figure on top is the original feature, below is the **fourth root** feature. By comparing two distributions, we can see that the origin feature has very closed data value among the total observations, which we think will not reflect the difference of observations and therefore contributes less in our logistic regression training process. So we choose a function, which can make the data distribution more dispersed, and apply it on the feature to create a new one



(a)

Figure 2: new feature

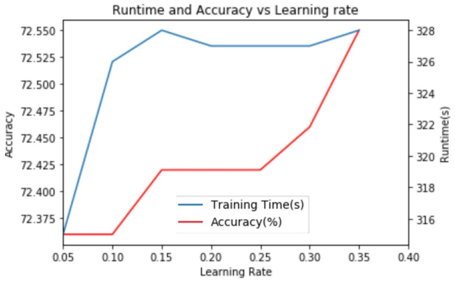
2.4 Ethical concerns

While working on publicly available patient disease dataset, it is hard to say whether the data contains certain private information of the patients. It is possible that some datasets may involve patients' names or addresses. Despite the fact that the dataset is public, downloading and working on it for machine learning purpose but without notifying the patients could be considered as an invasion of privacy.

3 Results

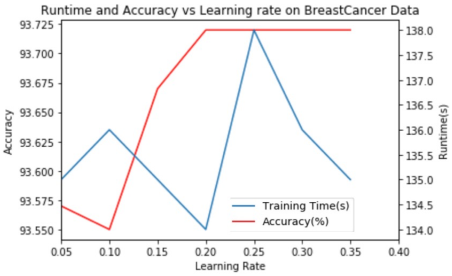
3.1 Logistic regression performance

For logistic regression approach performance, based on our figure for the wine quality dataset, we observed a slightly increasing trend on convergence speed as the learning rate increased. Increasing the learning rate also leads to the growth of accuracy until the model reached a limit accuracy at 0.3. When it comes to breast cancer dataset, the runtime maintained at the same level with minor fluctuations across the learning rates while the trend for accuracy conserves with the wine quality dataset.



(a)

Figure 3: Wine quality dataset performance



(a)

Figure 4: Breast cancer dataset performance

3.2 Runtime & Accuracy comparison

Note: This table is based on learning rate 0.3, which is our best performance learning rate.

	LDA	Logistic regression
runtime(s)	<0.15	400
accuracy(wine)	71.794%	72.671%
accuracy(breast cancer)	95.468%	93.428%

Table 1: Comparison

Obviously our table indicates that LDA approach behaves much better than logistic regression approach on runtime, and achieved similar accuracy. As logistic regression approach involves iteration, longer runtime was reasonable.

3.3 Feature subset and new features

With our subset of feature and new additional feature, the prediction accuracy increased by roughly 0.376% and 0.564%. In conclusion, additional feature **"fourth root"** apparently improved the performance of logistic regression model.

Logistic regression	original feature	feature subset
accuracy(wine)	72.546%	72.922%

Table 2: Feature Comparison

Logistic regression	original feature	dataset new feature
accuracy(wine)	72.546%	73.110%

Table 3: Feature Comparison

4 Discussion and Conclusion

Through this project, we explored the differences between using logistic regression model and LDA model. We also explored how features distributed and features transformation can affect on our model performance.

First, no matter which model we choose, understanding the performance of features is essential. Plotting the distribution of one single feature and dependence of two features can help a lot. We can further deal with our input data with these visualizations and then improve the performance of our model.

Second, during the experiments for two models, we found that the logistic regression approach and LDA approach fit for different sizes of the dataset. When training a small dataset, LDA performed better, and its accuracy closed to that of logistic regression approach while LDA approach greatly saved the runtime. As for a larger dataset, for example a dataset containing a few thousands of observations, we need to take a trade-off between stability of accuracy and the runtime. Although LDA approach run fast enough, the accuracy fluctuates for different training set. While logistic regression approach has a more stable accuracy but it will take much more time to train. This is also a topic that worth for further research in the future.

During this project, we improved our logistic regression model's performance by creating new features or a subset of input dataset. However, we still find some flaws in our results, which can be some directions for future investigation. For example, we tried many different subsets to improve the performance of logistic regression approach model, but most of them did not affect our accuracy significantly. In the future, we may choose other methods to select the best subset, such as using R or MatLab to see more statistical behaviors of our features, iterating all subsets of the input and choose the best. Also to get a more accurate prediction, we may try k-fold cross validation by increasing the parameter k or applying weighted linear least squares to improve the accuracy.

5 Statements of Contributions

Yichao Yang: task 2 and task 3, report writing

Weige Qian: task 1 and task 2, report proofreading

Hongshuo Zhou: task 3, report writng

References

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