Comp 551 Mini Project 1

Ian Benlolo (260744397), Thomas Racine (260801246), Alessia Woolfe (260721870)

***Abstract*— This project compares two classification methods on sample data sets. We implemented both logistic regression with gradient descent and linear discriminant analysis (LDA) and used k-fold cross-validation to train and validate the models. The models were trained and validates using the Wine Quality (red wine)(Cortez et al., 2009) and the Wisconsin Breast Cancer Database (January 8, 1991). We found that logistic regression takes significantly longer to train than linear discriminant analysis, and both methods have similar accuracy. We also tested how different learning rates affect logistic regression, and found that the highest learning rate was 0.0025. We also tried implementing monotonic decreasing functions to lower our learning rate based on the number of iterations to improve accuracy. We found that the various functions we tested marginally affected our result, but actually lowered the accuracy of our models. Finally, we preformed various experiments on our implementations to attempt to improve their accuracy such as feature selection and implementing Lasso Regression in our Logistic Regression model. These experiments failed to make a significant improvement in our models.**

I. Introduction

Logistic regression and linear discriminant analysis are forms of discriminative learning in which we estimate the probability of a given entry belonging to a class and output the class with the highest probability. Formally, logistic regression is based on linear regression where the relationship between the outcome and features is modelled with a linear equation, i.e. $y = w_0 + w_1x_1 + ... + w_mx_m$, where y is the output, w_i are the weights, and x_i are the given data points. This is used in a logistic function, $P(Y = 1 \mid x) = \sigma(w^Tx) = \frac{1}{1+e^{-w^Tx}}$, to ensure the output is between 0 and 1. Our goal is to maximize the likelihood function by modifying the weights w. In reality, we maximize the log of the likelihood function by minimizing the negative log-likelihood, or cross-entropy loss, using gradient descent. Other work done on the same Wisconsin breast cancer data set found that using logistic regression yielded high accuracy (Sultana, Jilani, 2018).

Linear discriminant analysis makes the assumptions that each class follows a Gaussian distribution and has the same co-variance matrix. Using Bayes' Theorem, we can expand $\frac{P(y=1|x)}{P(y=0|x)}$ into $\frac{P(y=1|x)P(y=1)}{P(y=0|x)P(y=0)}$. Then, we compute the log-odds ratio of this, given an input x and predict its class based on whether the log-odds ratio is positive or negative. Predictions can also be made using Bayes' rule to compute $P(y = 1 \mid x) \& P(y = 0 \mid x)$ and return whichever has the highest probability. The linear discriminant function $\delta_y(x) = x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + log(P(y))$ can be used to estimate those probabilities as it returns a value proportional to $P(Y == y \mid x)$ Since LDA does not use gradient descent, we found it much faster than logistic regression.

The project task is to analyze the two datasets, red wine quality and breast cancer data, and use logistic regression and linear discriminant analysis to train a model to make predictions about the data. The red wine quality data set has 1599 entries and 11 features, while the breast cancer data set has 699 entries and 10 features. Each entry also includes an output feature: ranging from 1-10 for red wine, and either malignant or benign for breast cancer.

II. Datasets

Since this is a binary classification task, the output variable for wine was defined as "good" (1) using a threshold of 6 out of 10. The cancer data set's classes were set as 1 for malignant and 0 otherwise. Some entries had to be removed from the breast cancer data set as they had missing data, reducing the entries to 683. The data features were normalized to the range $[0, 1]$ and the individual cases shuffled randomly every run.

A. *Ethical Concerns*

A few ethical concerns arise when dealing with sensitive information such as breast cancer and wine data. The breast cancer data raises some issues because it is data collected from real people. It is paramount that the anonymity of these people is protected. As well, if this data are used as training sets for real-life classification of tumors, there is the possibility of classification errors due to over-fitting/over-usage of the data set. This can lead to the demise of a patient due to faulty algorithms which is unacceptable. For the wine dataset, a concern may be a breach of patent/secret of a specific brand.

A first look at the two data sets hinted to what we'd expect as results.

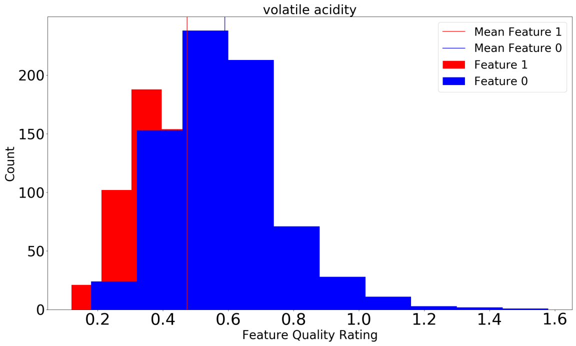


Fig. 1: The "good" and "bad" labeled data for Volatile acidity feature for wine data. This shows how close the mean of both distributions are, which is a common case for the wine dataset.

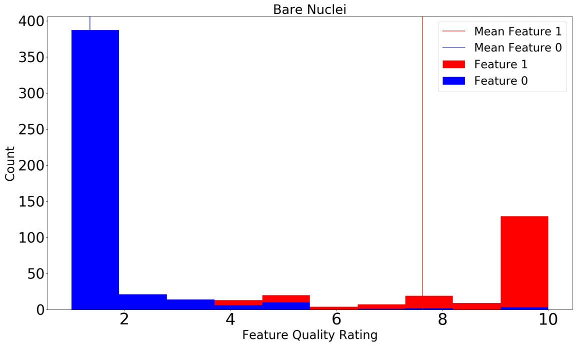


Fig. 2: Distributions for the malignant class (red) and the benign class (blue). The large disparity in their mean is a general case for the features in this data set.

See appendix-fig 18 and 16 for the other features. It is quite obvious from these plots that the cancer features have a larger difference between the mean of both classifications that in the wine data.

III. Results

A. *Effect of Learning Rate on Logistic Regression*

Learning rate in logistic regression had a more noticeable effect on the wine data set compared to the breast cancer data set. For wines, if the learning rate was too low, < 0.0001, or too high, > 0.1, then the accuracy suffered. We found that the best accuracy was when we chose a learning rate of 0.002, found by running cross-validation, iterating by 0.001 sized-steps. For the breast cancer data, the learning rate did not have a noticeable effect on the accuracy - nor was there a significant correlation between learning rate and accuracy. The graphs below compare different learning rates and the accuracy obtained with the specified learning rate for each data set using logistic regression.

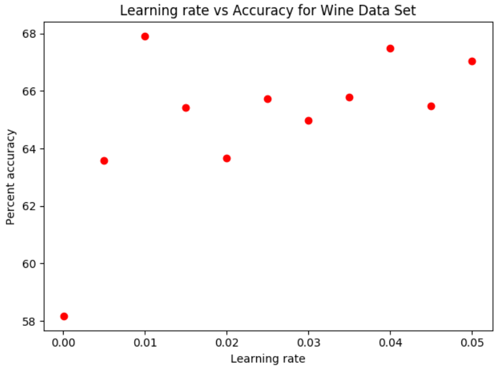


Fig. 3: The percent accuracy for the wine data set is plotted against the learning rate. There was no significant correlation between the variables.

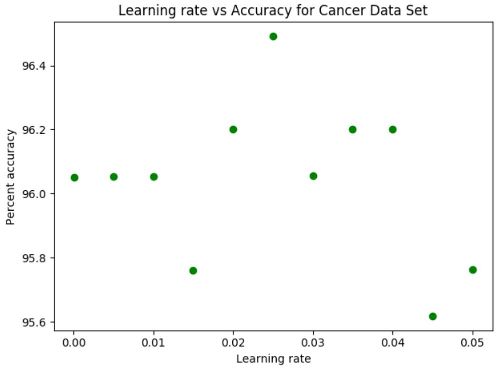


Fig. 4: The percent accuracy for the cancer data set plotted against the learning rate. Again, no significant correlation between learning rate and accuracy.

B. *Run-time and Accuracy of LDA vs Logistic Regression*

We compared linear discriminant analysis on the two data sets with logistic regression (1000 iterations, 0.0025 learning rate, and 0.5 lamba value for L1 regularization) and found that logistic regression performed slightly better on the breast cancer data, but slightly worse on the wines data. In each case, logistic regression was much slower than LDA. Table 1 outlines the run-time and accuracy for each data set.

	LDA	Logistic Regression
Runtime Wine (s)	0.07519	7.5357
Accuracy Wine (%)	73.108	72.920
Runtime Cancer (s)	0.02579	2.22031
Accuracy Cancer (%)	95.908	96.639

TABLE I: Runtime and accuracy of models

C. *Convergence with varying learning rate modifiers*

Functions were used to tune the learning rate at every iteration in order to converge more quickly and not "jump" around at local minima. Two of these were implemented and the results of both overlapping can be seen in Figure 5.

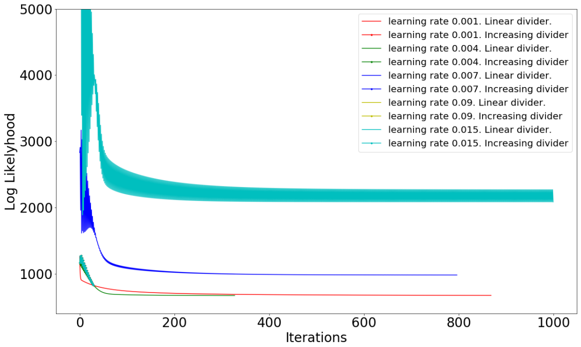


Fig. 5: A linear divider and a monotonically increasing on for the learning rate. This was run with a threshold of 0.01 for the δLog-Likelihood.

D. *Performance Improvement with Features*

The Pearson correlation matrix was computed for the wines data set in order to select a good subset of features for logistic regression. It is noted that in logistic regression, correlated features can cause inaccuracies (Yusuff, Mohamad, Ngah, Yahaya, 2012). Terms that had a larger than 60% correlation were removed. Fixed acidity was correlated to citric acid (67.17%), density (66.80%) and ph (-68.3%) and was removed, and free sulfur dioxide was correlated with total sulfur dioxide (66.77%) and was removed as well. Though we expected this to slightly increase the accuracy of our model's predictions it turns out it actually lowered it consistently by around $1-2\%$ for both models.

IV. Further Investigation

A. *Lasso regularization*

We studied the effects of L1 regularization on the models and found that it did not improve the results by much. We tested values ranging from 0.0001 to 100 and only saw results near the average accuracy of Logistic Regression on the wine dataset at very small λs which can be seen in Figure 6.

B. *Different Implementations of LDA*

To improve the accuracy of LDA, we attempted both finding the log-odds ratio as mentioned above, as well as finding the maximum of $\delta_y = x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu^T\Sigma^{-1}\mu_y + \log(P(y))$, the linear discriminant function proportional to $P(Y = y)$. This second implementation is found in Linear_discriminant_analysisBayes.py. The second implementation was slightly faster, with a runtime of 0.0620s versus 0.09146s on the wine data set, and identical accuracy of 73.1%. As well, we tried using both the built-in numPy function for finding the covariance matrix and our own implementation - for wines, that was 73.1% accuracy with our own implementation versus 73.4% with numPy's built-in cov() function.

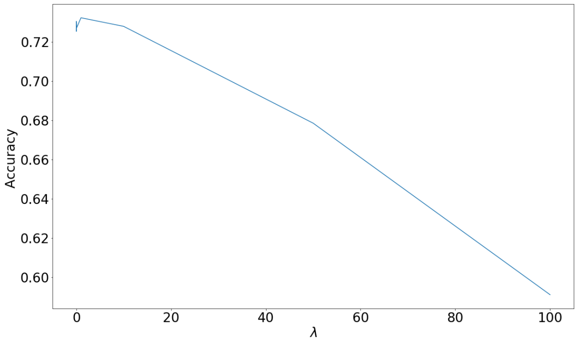


Fig. 6: The accuracy of Logistic Regression on the wine dataset with varying L2 regularization constants λ.

V. Discussion and Conclusion

The various optimizations that were discussed and tested did not seem to improve the accuracy of our logistic regression model (tested on the wine dataset). This is thought to mainly be attributed to the simplicity of our model. Over-complicating a simple model can (and has, as shown) even lead to lowering the accuracy of predictions. Regularization is used to reduce variance but introduce some bias. We have to be careful when using regularization in a simple model in order to avoid this bias. Furthermore, we were working with relatively small datasets which may be another factor contributing to these failings.

To do a more in-depth analysis of these models and to find better ways to improve the accuracy of our models, a interesting experiment would be to try higher order fits for our logistic regression model. L2 Ridge regression could also be implemented and a mix of both regularization method could be tested.

Since Regularization can be used in any model, adding some sort of regularization to LDA may lead to an improvement in accuracy.

VI. Statement of contributions

Ian worked on implementing Logistic Regression, Alessia worked on LDA and Thomas took part in helping both with their implementation. The preprocessing was done all together and so was the writeup of the report.

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. ISSN: 0167-9236.

[2] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 18.

[3] Sultana, J. (2018). Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers. International Journal of Engineering and Technology(UAE).

[4] Yusuff, H., Mohamad, N.S., Ngah, U.K., Yahaya, A.S. (2012). BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION.

Appendix

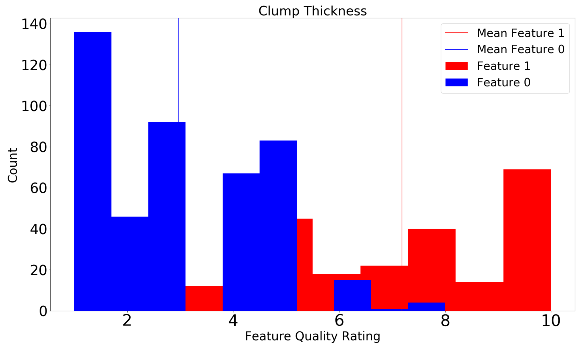


Fig. 7: 1a

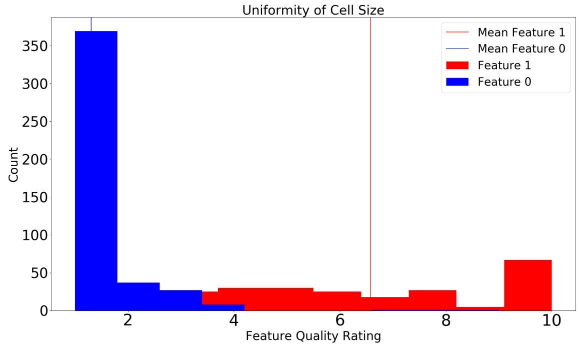


Fig. 8: 1b

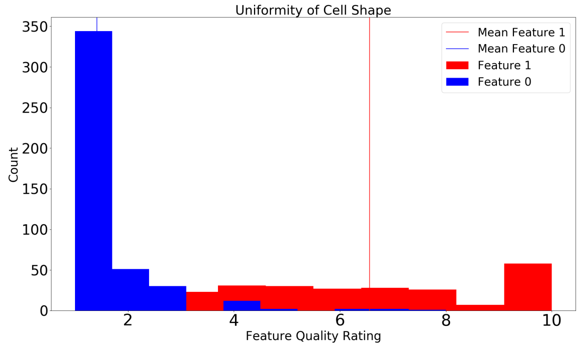


Fig. 9: 1a

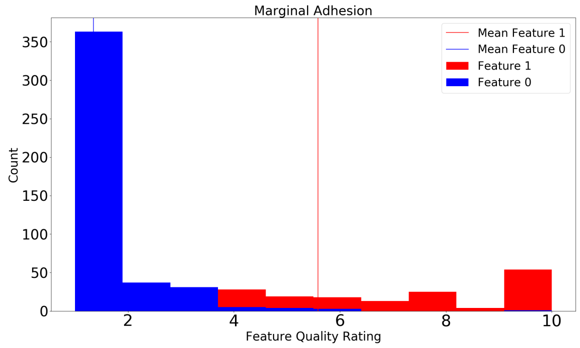


Fig. 10: 1b

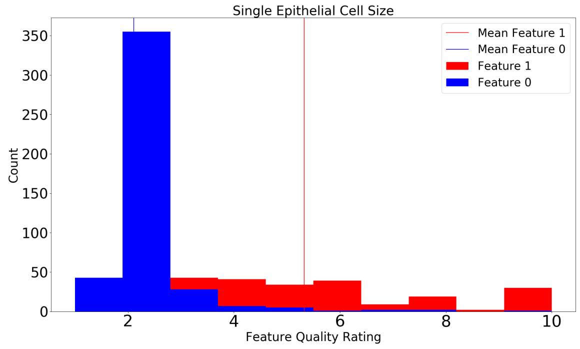


Fig. 11: 1a

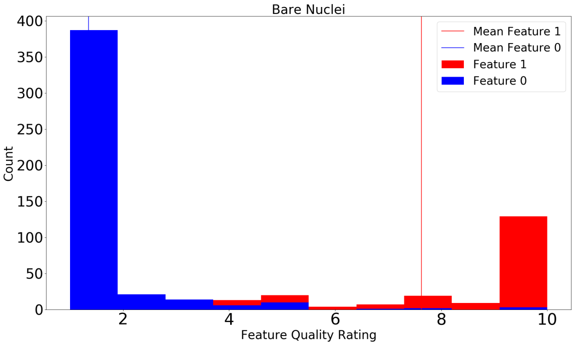


Fig. 12: 1b

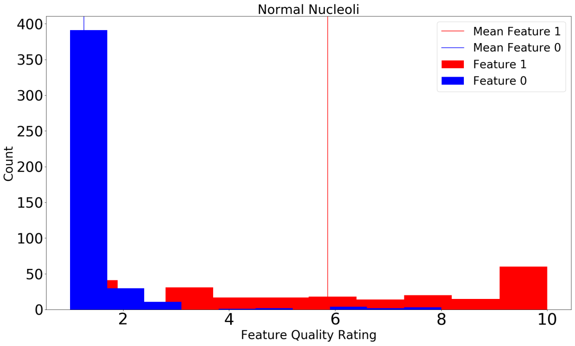


Fig. 14: 1b

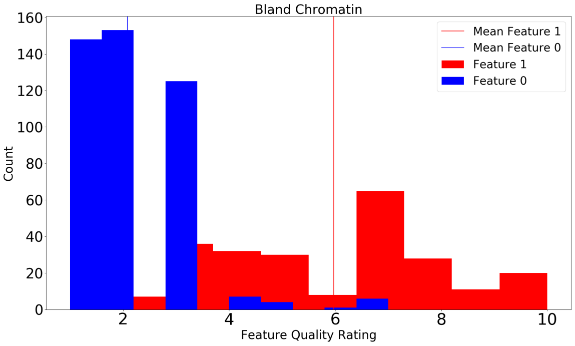


Fig. 13: 1a

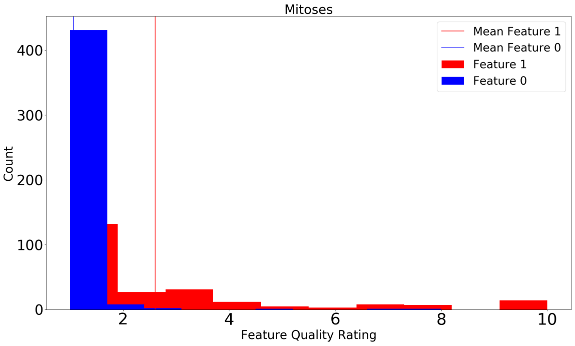


Fig. 15: 1a

Fig. 16: Plots for the cancer data showing the disparity between both classifications for each feature.

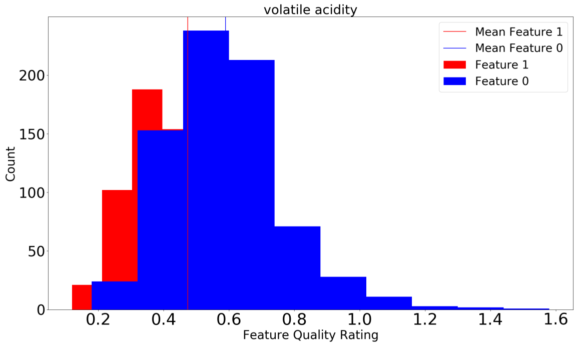
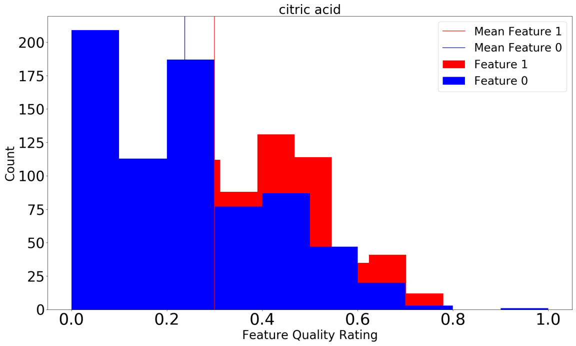
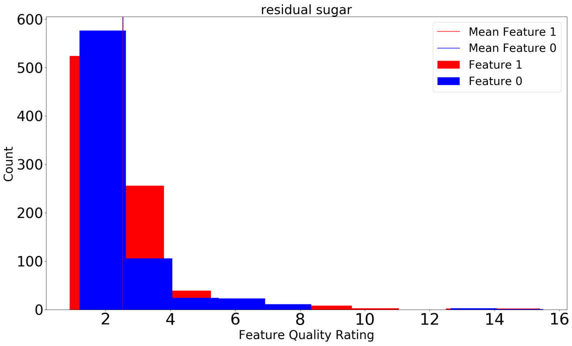
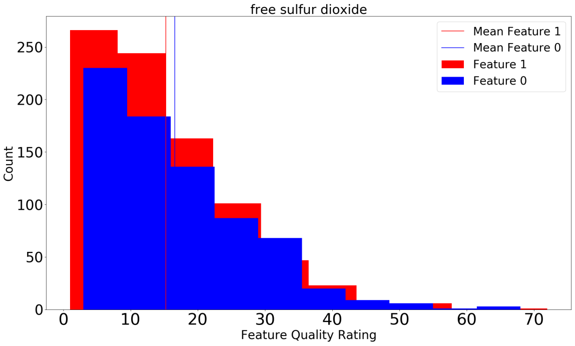
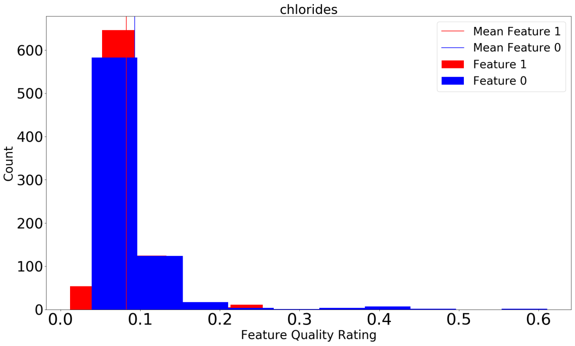


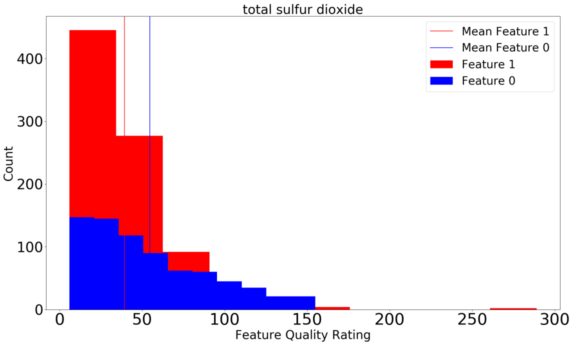
Fig. 17: 1a




residual sugar
Mean Feature 1
Mean Feature 0
Feature 1
Feature 0
Count
0
100
200
300
400
500
600
2
4
6
8
10
12
14
16
Feature Quality Rating


free sulfur dioxide
Mean Feature 1
Mean Feature 0
Feature 1
Feature 0
Count
0
50
100
150
200
250
0
10
20
30
40
50
60
70
Feature Quality Rating


chlorides
Mean Feature 1
Mean Feature 0
Feature 1
Feature 0
Count
0
100
200
300
400
500
600
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Feature Quality Rating


total sulfur dioxide
Mean Feature 1
Mean Feature 0
Feature 1
Feature 0
Count
0
100
200
300
400
0
50
100
150
200
250
300
Feature Quality Rating

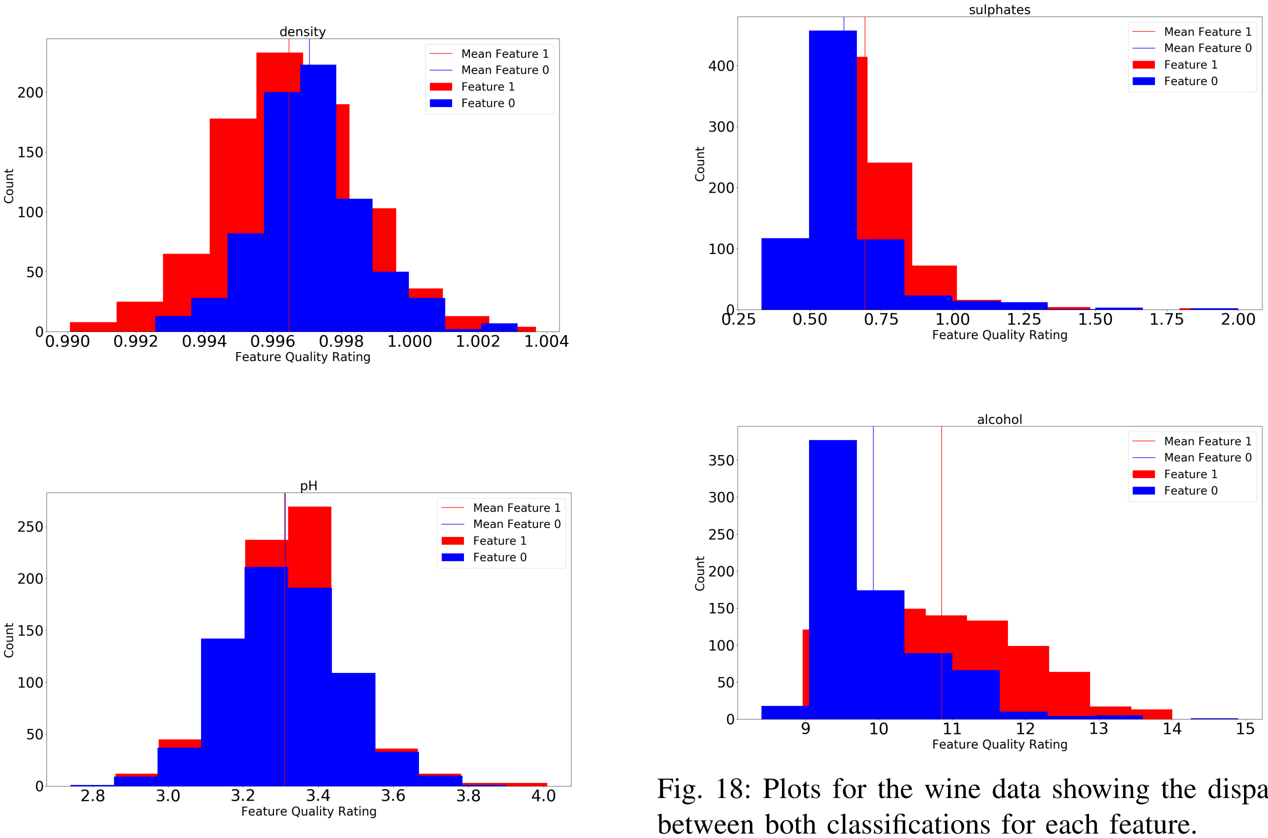


Fig. 18: Plots for the wine data showing the disparity between both classifications for each feature.