COMP 551: Applied Machine Learning

MiniProject 1 - Group 76

Team Members:

Chris Axon - 260735894
Michel Kassis - 260662779
Evan Savage - 260831481

Abstract

The objective of the project was to explore different linear classification models, logistic regression (LR) and linear discriminant analysis (LDA). More specifically, to gain experience implementing the models from scratch and comparing their performances on two distinct benchmark datasets. The first dataset's classification goal was to predict the taste quality of wine based on chemical measurements. The second dataset's classification goal was to predict whether a tumour is malignant or benign based on various physical properties. In our results, we found that the LDA model was significantly faster to train than LR. In respect to accuracy, the LDA approach consistently yielded better results than LR, even through repeated experimentation and tweaking of LR parameters.

Abbreviations and Notations

ML: Machine Learning
LR: Logistic Regression
LDA: Linear Discriminant Analysis

Introduction

For this project, the team was assigned with implementing logistic regression (LR) and linear discriminant analysis (LDA) from scratch without the assistance of available machine learning libraries for python. By tweaking both required and self-determined parameters of both models, we worked to achieve the best possible accuracy on two separate datasets. There has been research papers comparing both methods of classification, specifically S. James Press & Sandra Wilson (1978)'s paper which compares them in depth and reported their own supportive empirical studies.

Two datasets were used in this project: a wine quality and breast cancer classification dataset. The wine dataset is composed of chemical features that contribute to a wine quality on the output. The breast cancer dataset includes various physical features of tumors that lead to a benign or malignant classification.

Using a k-fold cross validation, we implemented the two algorithms with an added regularization feature extension. The wine dataset proved to be trickier to train than the breast cancer dataset, resulting in much lower classification accuracies across the board. The ridge regression regularization greatly helped the classification accuracy for the wine dataset, while minimally influencing the outcome on the cancer dataset.

Datasets

Data Source & Size

The wine dataset is sourced from the UC Irvine Machine Learning Repository. It compares different variants of the Vinho Verde wine. There are eleven features which can be chemically measured, including acidity, sugar, and concentration of different compounds, where each wine receives a "quality" score between zero and ten. There were 1599 distinct wines tasted.

The breast cancer dataset is also from the UC Irvine Machine Learning Repository. It contains nine features, including radius, texture, and perimeter, calculated from digital image of breast cancer cells and whether the cell is classified as benign or malignant. There were 683 complete examples.

Data Cleanup

Both datasets were initially loaded from their respective text files using the numpy library in python. Both datasets were then run through a cleaning phase, which removed examples that contained any non-numeric entries. The first column of the breast cancer dataset was always excluded from processing, since it contained unique identifier values, which have no inherent effect on the output classification.

Decision Boundary for Datasets

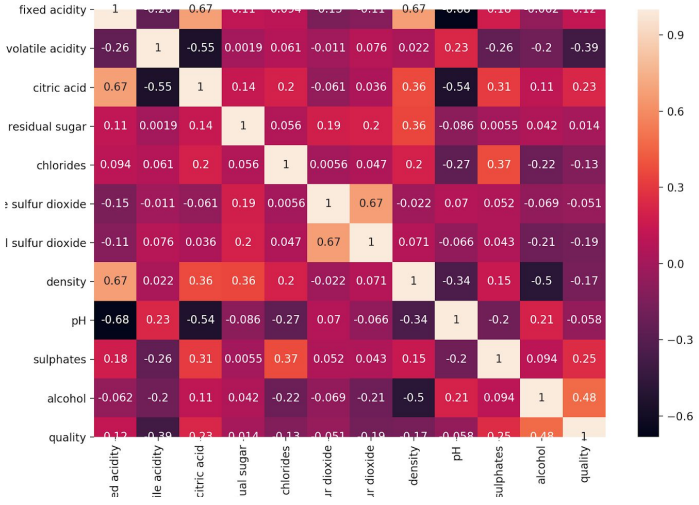
As this assignment is concerned with binary classifications, the output variable in both datasets had to be thresholded so that points above a certain value would result in a classification of 1, and 0 otherwise. For the wine dataset, the accompanying "winequality-red.names" file indicated that the wines were rated on a scale from 0-10, and a threshold value was chosen to be 5 (equal to five is classified as 0). For the breast cancer dataset, the benign (2)/malignant (4) output was already a binary task, so it was trivial to threshold the output at 3, though it made the column numerically compatible for binary classification processing.

Ethical Concerns

In any project, there's always ethical concerns that have to be considered while in development. However, there's extra concern in particular for developing a binary classifier for breast cancer diagnosis as it has as much power to be a misinformative and potentially harmful as it equally has power being informative. In most ML projects, we have to choose if we would rather have a false positives versus false negatives. For example, a false positive diagnosis would require someone to consider treatment who does not have cancer, though it is good to be on the safe side from an applied perspective. Even worse, a false negative diagnosis would leave someone living with the condition without treatment.

Visualization

To investigate our datasets further, we plotted a heatmap that showcases all the features of a given dataset and represents their correlation strength to each other. This allowed us to understand our datasets better in terms of each features significance to the output. This proved to be useful, especially for the wine dataset as it has a lot of non useful features.It also provides us which features contribute to quality(e.g Alcohol & Quality = 0.48), as shown in figure adjacent.



New features and figures

Regularization

Regularization is a form of regression that shrinks the coefficients estimates according to their significance to avoid overfitting the model by increasing the bias and decreasing the variance. There are many regularization techniques including Ridge, Lasso and Elastic Regularizations. LR using gradient descent can benefit from regularization to decrease the weights that could

potentially cause overfitting as discussed in Cessie, S. and Houwelingen, J. C. (1992), Ridge Estimators in Logistic Regression.

Options for ridge and lasso regression regularization were added to LR based on an additional input parameters, regularization type and lambda. Lambda is the regularization coefficient, scaling the penalties from the regularization term. Ridge regression is used to bolster feature weights by imposing a penalty on their respective sizes. As opposed to lasso regression, it does not completely suppress certain features; rather, it reduces their respective weights across the board. Lasso regression will tend to move feature weights to zero that are not influencing the output. Elastic regularization is a hybridization of both ridge and lasso methods, and it is also implemented in the code for LR.

Results

Wine	Parameters	T (s)	Acc	Cancer	Parameters	T (s)	Acc
LR	Iter: 2500 | Rate: 0.008	116.4	58.10	**LR**	Iter: 1200 | Rate: 0.5	28.8	94.88
LR (Ridge)	Iter: 2500 | Rate: 0.009 Lambda:0.1	116.8	71.86	**LR (Ridge)**	Iter: 1100 | Rate: 0.55 Lambda: 0.01	28.69	96.19
LR (Lasso)	Iter: 2500 | Rate: 0.009 Lambda: 0.1	114.9	57.16	**LR (Lasso)**	Iter: 1100 | Rate: 0.55 Lambda: 0.01	25.3	95.17
LR (Elastic)	Iter: 2500 | Rate: 0.009 Lambda: 0.1	111.8	72.30	**LR (Elastic)**	Iter: 1100 | Rate: 0.55 Lambda: 0.01	25.2	96.34
LDA	N/A	0.070	61.29	**LDA**	N/A	0.030	94.73

Logistic regression performance is dependent upon the interaction of iterations and learning rate. Learning rate controls the step size that the gradient descent algorithm moves towards the optimal weights for the model. If it is too large, the weights can continually oscillate around optimal weights without ever converging, and if it is too small, the model may never approach its optimal weights. Through repeated experimentation, the learning rates mentioned in the table above were settled upon.

In regards to runtime, the LDA algorithm performs consistently faster than the runtime for LR. The runtime for LR is affected as a function of the input iterations, whereas the runtime of LDA is consistent since it takes place irrespective of specified iterations. LDA was also consistently more accurate than vanilla LR, since it did not involve the tweaking of input parameters to increase accuracy to a certain point.

Once finding optimal input parameters for vanilla LR, the accuracy for the wine dataset LR almost reached the accuracy for LDA. For the breast cancer dataset, the accuracy was similar between vanilla LR and LDA once settling on veritable inputs.

Ridge regression greatly improved the accuracy for the wine dataset, by about 14%, whereas lasso regression had little to no effect on the accuracy for both datasets. Elastic regression tended towards the results for independent ridge regression.

Discussion and Conclusion

In this project, we have compared the performance of two methods of classification, logistic regression using gradient descent and linear discriminant analysis. We've measured their accuracies on two distinct datasets, wine dataset in which we predicted if a wine is good or not based on its chemical properties, and a breast cancer diagnosis dataset in which we predicted if a tumour was malignant or benign based on other properties. We determined that the basic LR algorithm does comparably to the basic algorithm LDA. However, improving the LR algorithm with rasso and/or ridge regression can significantly increase the rate of accuracy of logistic regression. Both algorithms performed significantly better than the naive *most_frequent* algorithm (53% accuracy for the wine data and 65% accuracy for the cancer data), and would be a very useful tool to aid in patient diagnostic, in addition to physician screening.

We have also gained experience developing a binary classifier from scratch and learning how to improve its accuracy. This includes data preprocessing, cleaning up the data, implementing k fold to split the data into training and testing data. Afterwards, we worked on fitting the models, predicting the labels and measuring each method's accuracy with and without regularization. We also decided to implement different regularization techniques to figure out if our model was suffering from overfitting. Implementing them also allowed us to experiment on using different penalty coefficients for ridge, lasso and elastic regression which gave us various accuracies, so we had to decide which penalty factors worked with each model and dataset.

Future investigations would include changing our features even further to check how much higher accuracy we can achieve. We can also test our models on different datasets with different number of features to compare even further how much the accuracies change for LDA and LR when we test on datasets with more features. We can also learn to develop a multi class logistic regression model so it would categorize wine datasets further. Instead of binary classifying them into good or bad only, we can categorize them into different flavours such as fruity/sugary/sour according to their properties.

Statement of Contributions

The work on this project was shared equally among all three team members. Evan implemented the LR and LDA algorithms including regularization, Chris handled the k-fold cross-validation and modularization of code, and Michel assisted in both prior implementations, developed data visualizations and suggested data cleaning and interaction. All three team members contributed to their respective areas of the report.

References

S. James Press & Sandra Wilson (1978) Choosing between Logistic Regression and Discriminant Analysis, Journal of the American Statistical Association, 73:364, 699-705, DOI: 10.1080/01621459.1978.10480080

Cessie, S. and Houwelingen, J. C. (1992), Ridge Estimators in Logistic Regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41: 191-201. doi:10.2307/2347628