Machine Learning - Project 1

Boyer Eymeric, Lagergren Hugo, Mahmoud Fouad
School of Computer Science, EPF Lausanne, Switzerland
28 October 2023

I. Introduction

The purpose of this project is to predict coronary heart diseases using data of more than 300'000 individuals from the dataset gathered by behavioral Risk Factor Surveillance System.

II. Methodology

First, we analyze the data to determine the necessary pre-processing steps. In this project, we employed four algorithms: Least Square, Ridge Regression, Logistic Regression, and Regularized Logistic Regression. We utilize 10-fold Cross Validation to evaluate and fine-tune the models, ensuring we don't over-fit to a particular segment of the training data. Although we measure outcomes using precision, recall, and accuracy, only accuracy figures and F1 scores are presented in this paper. This is because our test set predictions are assessed on this metric on AI crowd. Following the various optimization processes, we choose the most optimal model.

III. Feature Selection and Data Pre-processing

An initial data examination is carried out to gain a preliminary understanding of the dataset. We observe that the 26 first variables don't contain any useful information about each participant and thus can be dropped. For the Data-preprocesing, the first step is to remove the outliers. We use the Tukey Fences criterion for Outliers and replace the values with the mean. We then look for features with more than 30 percent of NAN values and drop them. It is then time to replace the remaining nan with the mean aggragated over all samples and then to normalize the data. Features with standard deviation being 0 are removed as they don't bring any information to help the binary classifications. They also are a lot of variables that are badly encoded such as for the question "how often do you have a shower a week?". The number 8 was picked for the answer 0. Similar cases where found where the number 7, 9, 77, 99 meant that the person didn't answer the question. This could partially explain the "poor" results.

We have the scores with the outliers in table 1 and then without them in table 2. As expected there is a clear gain in performance as those values which were causing some overfitting our models are discarded.

	iterations	γ	λ	Accuracy(%)	F1 score
Ridge Regression	-	-	10^{-3}	64.4 ± 0.1	0.304
Least Squares	-	-	-	67 ± 0.1	0.181
Log Regression	10^2	0.1	-	64.6 ± 0.1	0.304
Reg Log Regression	10^3	0.5	10^{-4}	68.1 ± 0.1	0.321

TABLE I: scores with outliers

	iterations	γ	λ	Accuracy(%)	F1 score
Ridge Regression	-	-	10^{-3}	66.9 ± 0.1	0.318
Least Squares	-	-	-	66.7 ± 0.1	0.317
Log Regression	10^2	0.1	-	67.6 ± 0.1	0.320
Reg Log Regression	10^3	0.5	10^{-4}	70 ± 0.1	0.333

TABLE II: scores without outliers

Outliers can have a pronounced influence on regression models, particularly those that rely on the minimization of squared errors, such as Least Squares. We can clearly see this in the graph as the F1 score for least squares is far lower than the others but catches up when the outliers are removed. Choosing the right parameter for each model can greatly affect obtained results as can be seen in the table below for the ridge regression model. We observe that the smaller the lambda parameter is, the higher our accuracy and total score, which is defined as being the sum of our F1 and accuracy scores.

Lambda value	Loss	F1 Score	Accuracy(%)	Total Score
0.00001	0.475	0.301	0.638	0.939
0.0001	0.475	0.299	0.634	0.933
0.001	0.476	0.297	0.632	0.930
0.01	0.476	0.298	0.631	0.929
0.1	0.476	0.297	0.629	0.926

TABLE III: Evaluated metrics per lambda value

IV. Best Model

From our pre-processing outcomes, it's evident that the cross-validation (CV) accuracy and F1 score can vary (improving or deteriorating) based on the specific pre-processing method employed. There isn't a one-size-fits-all solution to enhance all models simultaneously. We tried various methods such as removing thoses features with low variance, small pearson correlation and and fined tuned our parameters. We get our best scores by setting the threshold to 0.23 with the regularized logistic regression.

	iterations	γ	λ	Accuracy(%)	F1 score
Ridge Regression	-	-	10^{-3}	85.5 ± 0.1	0.416
Least Squares	-	-	-	85.3 ± 0.1	0.414
Log Regression	10^3	0.1	-	76 ± 0.1	0.361
Reg Log Regression	10^3	0.5	10^{-4}	75 ± 0.1	0.355

TABLE IV: Final scores

V. Conclusion

In summary, we've underscored the significance of data pre-processing. Techniques like least squares, for instance, only yielded effective results after several pre-processing measures were applied.