Machine Learning : Project 1 report

Max Tost, Jéremy Salm, Gauthier Leurent
EPFL, Lausanne, Switzerland
max.tost@epfl.ch jeremy.salm@epfl.ch gauthier.leurent@epfl.ch

November 1, 2024

***Abstract*—This report analyzes data from around 300,000 individuals regarding their lifestyle factors to employ machine learning techniques for evaluating the risk of developing cardiovascular disease. k-fold cross-validation is used, where weights and losses are computed using regularized logistic regression. The consequence of hyperparameter tuning on the regularization parameter has been studied. The double descent phenomenon is observed in test errors, highlighting the importance of regularization and extensive data cleaning. Over the 109,379 people in the test group, 1049 were predicted to contract a cardiovascular disease.**

I. Introduction

This work aims to determine the risk of a person developing Cardiovascular Diseases based on features of their personal lifestyle factors. This is indeed of great interest, since according to World Health Organization, Cardiovascular Diseases are becoming one of the leading causes of death globally [1]. The data comes from a system of telephone surveys focused on health of U.S. residents, Behavioral Risk Factor Surveillance System [2]. 328,125 people took part of the survey, answering 321 questions. The dataset representing all respondents, their answer to each question being features, has been first processed by filling missing answers, removing irrelevant features such as identifiers (Ids), and normalized. Then k-fold cross-validation has been performed using k=5, using regularized logistic regression to compute weights and losses. The optimal regularization hyperparameter λ has been determined by taking the best from the cross validation. We observed a double descent pattern in test loss across iterations, suggesting that increasing iterations in optimization can reduce overfitting.

II. Methods

In the dataset, respondents were classified as having coronary heart disease, labeled by a $y = 1$, if they have been told by a provider they hade coronary heart disease, or if they have been told they had a heart attack or angina. Respondents are labeled with a $y = -1$ if they do not have coronary heart disease. If no answer is given by a respondent to a question, the label is then $y = NaN$.

Data has first been loaded, then processed. Cross-validation has been performed, using regularized logistic regression with $y \in \{0, 1\}$. the optimal weights w_* have been computed using

$$w_* = \frac{1}{N} \sum_{n=1}^{N} -y_n x_n^\top w + \log\left(1 + e^{x_n^\top w}\right), \qquad (1)$$

where x_n is the data associated to the n^{th} individual, y_n is its label, and w are the weights [3].

A. Data Loading and Preprocessing

The data loading and preprocessing has been made using `run.py`. It includes the following functions :

- `load_csv_data()` loads data from CSV files, converting features X and binary labels y (from -1, 1 to 0, 1).
- `clean_and_standardize()` preprocesses features by normalizing them, imputing missing values with the column mean, and adding a bias term to each data point. Normalization ensured that features were scaled, in order to improve gradient-based optimization.
- During preprocessing, irrelevant features like identifiers and time stamps were removed. Initial tests showed that including identifiers led to disproportionately high weights for these features, likely due to non-informative data patterns that interfered with learning.

This cleaning process is essential for reproducibility, and it ensures that the model learns from meaningful patterns in the data.

B. Cross-Validation and Hyperparameter Tuning

The effect of regularization has been assessed using k-fold cross-validation, implemented in `cross_validation_reg_log()` to evaluate a range of regularization parameters λ. This setup is particularly effective for generalization testing and model stability.

The functions used in this process are :

- `build_k_indices(y, k_fold, seed)` generates indices for k-fold cross-validation, where $k = 5$ was selected to balance training and validation set sizes.
- For each λ value, `cross_validation()` iterates over k-folds, training on $k - 1$ folds and validating on the remaining fold. This function integrates `reg_logistic_regression()` which optimizes model weights for regularized logistic regression.

The cross-validation process calculates average training and test losses across folds, with the optimal $\lambda = 1.468 \cdot 10^{-6}$ identified based on the lowest test error.

III. Results

The results obtained in this work are now described: first a brief illustration of the predictions regarding the test group is shown. Secondly, the study highlighting the impact of regularization of test and training errors is performed.

A. Repartition of predicted individuals for the test group

Quantitatively, the predictions performed on the test group, composed of 109,379 individuals, are described in Tab.I. It can be deduced that the majority (99.04 %) of them are predicted to not contract a cardiovascular disease (-1), while the rest (0.96 %) are predicted to contact one (+1).

Type of prediction	**number of people**	**% of test group**
+1	1049	0.96
-1	108330	99.04

Table I: Summary of the repartition between individuals that are predicted to contract cardiovascular disease (+1) or not (-1) according to the model discussed in this study

B. Effect of Regularization on Test and Training Error

The test and training losses were computed for a certain range of λ. A first study was performed with 300 iterations and a learning rate of 0.1 and a second one was also conducted but this time with 1000 iterations and a learning rate of 1. These results are presented in Fig.1 and Fig.2, respectively.

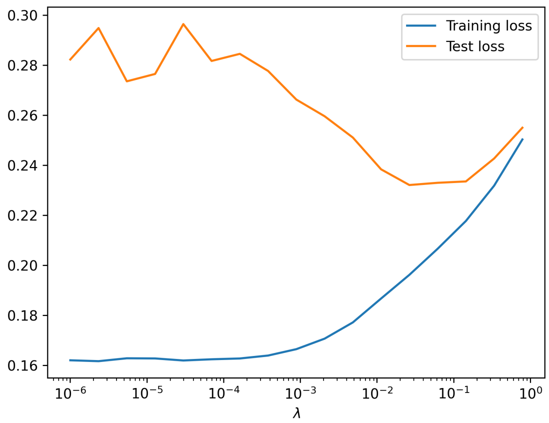

Figure 1: Computing training (in blue) and test (in orange) losses, with 300 iterations and learning rate 0.1

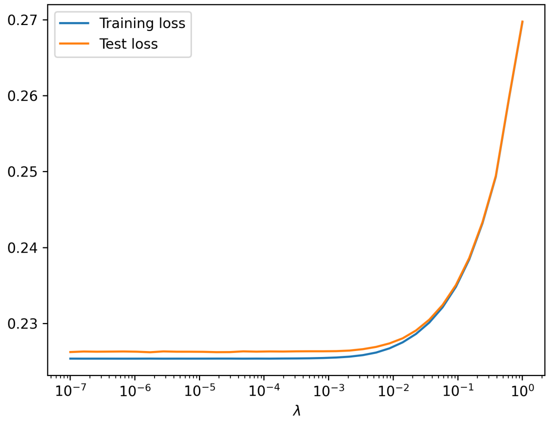

Figure 2: Computing training (in blue) and test (in orange) losses, with 1000 iterations and learning rate 1

As a first observation, the value of the training loss is always smaller than the value of the test loss, for all values of λ in both cases, which is expected. In addition, in the first case, it can be noticed that for values of λ between 10^{-6} and 10^{-3}, the difference between the two losses is large and stay constant on average. Then, from $\lambda = 10^{-3}$, this difference starts to decrease. In the second case, the difference between the two losses is also constant on average for a range $\lambda \in [10^{-7}, 10^{-3}]$, but its value is now much smaller. Then, for $\lambda > 10^{-3}$, this difference also decreases as the two losses are getting closer and closer.

IV. Discussion

Data cleaning is an important part of data preprocessing. Indeed, the standardization or the removal of non-significant features can affect positively the results quality. For example, removing the feature "Ids" of the input data, which are the individual number, will surely improve the model by avoiding the weights to be influenced by this feature.

Double descent [4], a recurring pattern in modern machine learning, was observed in our tests. Fig.1 shows that at low iteration counts, the test error has a minimum around 0.23 for a λ around 10^{-2}, suggesting optimal performance. However, as iterations increased as seen in Fig.2, test error for lower regularization values stabilized just under 0.23, becoming largely independent of λ. Hence here the increasing number of steps helped to reduce overfitting.
The double descent phenomenon, which can be identified in the current study, is in line with recent machine learning findings [5], where complex models benefit from extended training to achieve generalization stability. In our case, the results confirm that iterative tuning of regularization can help reach a stable error plateau, balancing both underfitting and overfitting.

V. Conclusion

To conclude, the work, based on data from Behavioral Risk Factor Surveillance System survey, has highlighted several aspects.

The data have been first loaded and processed. The importance of removing irrelevant features as "Ids", and processing has been shown (see Sect.II). Then, using cross validation and regularized logistic regression with the optimized parameter $\lambda = 1.468 \cdot 10^{-6}$ numerically found, predictions have been successfully performed. Individuals in the test group have been labeled to contract or not a cardiovascular disease (see Sect.III), resulting to be 0.96% and 99.04% of the test group, respectively. Also, double descent phenomenon in the optimization of the parameter λ has been observed (as discussed in Sect.IV).

References

[1] World Health Organization, Cardiovascular Disease, 11 June 2021 https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] CDC - 2015 BRFSS Survey Data and Documentation, last visit : Nov 1, 2024, https://www.cdc.gov/brfss/annual_data/annual_2015.html

[3] N.Flammarion, M.Jaggi, CS-433 Machine Learning lecture 5b, EPFL, 2024,https://github.com/epfml/ML_course/blob/main/lectures/05/lecture05b.pdf

[4] N.Flammarion, M.Jaggi, CS-433 Machine Learning lecture 4b, EPFL, 2024,https://github.com/epfml/ML_course/blob/main/lectures/04/lecture04b.pdf

[5] Z. Deng, A. Kammoun, C. Thrampoulidis, *A Model of Double Descent for High-Dimensional Logistic Regression*, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 4267-4271, doi: 10.1109/ICASSP40776.2020.9053524.