Faculty of Computer and Communication Sciences (IC)

Computer Science

CS-433 Machine Learning – Assignment 1

Authors

Ziyang He, Roméo Estezet, Capucine Denis

Assignment 1

October 2025

Abstract

We train and apply two machine learning models – regularised logistic regression and ridge linear regression to the dataset provided by AIcrowd. The aim is to compare the ability of the two models to correctly predict the likelihood of coronary heart disease (CHD) using the F1 score metric. We show that the logistic regression model performs much better than the linear regression model in this classification problem.

1 Methodology

We divide our method into two major processes: data preprocessing, model training. For model training, we will compare the results of regularised logistic regression with ridge regression. Let us state the main steps of each process below and elaborate.

1.1 Data preprocessing

The data preprocessing steps are given as follows:

1. Convert invalid entries into NaN values and suppress features by removing columns which are either completely irrelevant, or contain too many NaN values.
2. Classify the remaining columns into categorical features and continuous features. One-hot encode categorical data to binary valued vectors.
3. Impute remaining NaN values in categorical columns with mode imputation and the remaining NaN values in continuous columns with mean imputation, and recombine continuous and categorical features. The result of the preprocessed training data matrix should have no NaN values.

1.2 Model training

The model training steps for logistic regression are given below:

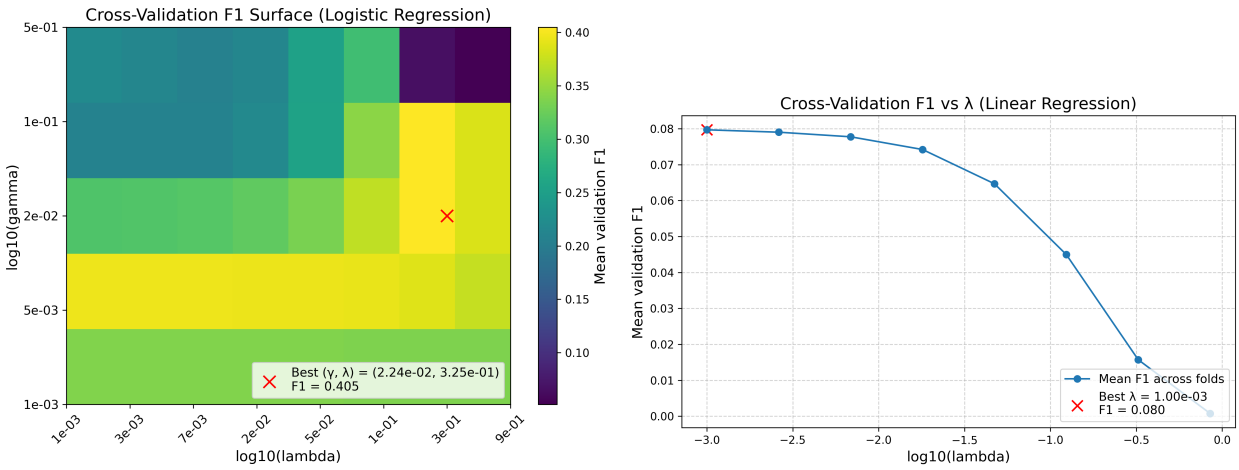
1. Split the preprocessed training data into $K = 5$ cross-validation folds, and set up a list of values for the step size γ and the regularised logistic regression variable λ. We use principal component analysis (PCA) to reduce the dimensionality of the training data to k principal components and optimise the model on (γ, λ).
2. For each (γ, λ) pair, for each training/validation split in cross validation, fit PCA on the training fold only; project both training and validation data onto the k principal component axes; train weights w using regularised logistic regression on training fold; evaluate the F1 score on validation fold.
3. Average the F1 score on each (γ, λ). Choose the (γ, λ) pair which has the highest F1 score.
4. Using this pair of (γ, λ), retrain the model on the whole preprocessed training dataset. This is our final model which will be applied on test data.

Note that training the ridge linear regression model involves exactly the same steps but with mean squared error loss instead of NLL, and we optimise only over λ due to the closed-form nature of linear regression (gradient descent not required). Also, before applying PCA transform, we always recenter the data to have a mean of 0 and standardise it by dividing by the standard deviation. To understand the importance of PCA and its mathematical relation to singular value decomposition (SVD) as implemented in the code, we direct the interested reader to a Chapter 12 of the following textbook by Kevin P. Murphy. Apart from this, we direct readers interested in understanding the F1 score (with values in $[0, 1]$, the higher the better) and why it matters to the following webpage.

The number of principal components k is chosen so that it represents enough of the dataset while having a low computational cost. Therefore we chose $k = 90$, i.e. the number of features in the preprocessed training data gets reduced from more than 200 to 90. When using regularised logistic regression, we also chose the max iterations to be 200 to balance accuracy with computational cost. When iterating over γ and/or λ, we set the range to be $[0.01, 0.5]$ and $[0.01, 0.85]$ respectively, in logspace base 10.

2 Model comparison

We first comment on the best pair of (γ, λ) chosen for the regularised logistic regression model and the best λ chosen for the ridge linear regression model. As seen in Figure 1, the logistic regression model obtains the best average F1 score of $F1 = 0.405$ for $(\gamma, \lambda) = (0.0224, 0.325)$ whereas the linear regression model obtains the best average F1 score of $F1 = 0.0797$ for $\lambda = 0.001$. For this reason alone, it is more appropriate to use the logistic regression model to classify this dataset, which is expected since unlike logistic regression, linear regression is not specialised for classification problems.



(a) Logistic regression model

(b) Linear regression model

Figure 1: F1 scores of regularised logistic regression model vs ridge linear regression model. Observe that the F1 score of the former is much higher than the latter, and the F1 score of the latter plateaus at 0.08 even when reducing λ to smaller values $< 10^{-3}$.

By submitting predictions of models on AIcrowd, the respective F1 scores on test data is shown in Table 1.

Table 1: F1 scores of trained models applied on test data

Model	**F1 Score**
Logistic Regression	0.412
Linear Regression	0.084

We observe that much like the training data, the logistic regression model performs much better than the linear regression model on test data when measured using the F1 metric. We conclude that the logistic regression model is preferred over the linear regression model in such classification problems.

3 Evaluation and improvements

We list and explain key points which could have further improved the result of our investigation.

- When preprocessing training data, we only manipulated features of data points and did not manipulate amount of data points. Thus the training dataset had much more -1 than $+1$. This meant model predictions likely favoured the majority class of "CHD negative" rather than the minority class of "CHD positive", which is not good for such crucial disease prevention purposes.
- We applied PCA to reduce the dimensionality of the dataset and facilitate training. However, with majority features in training dataset being categorical instead of continuous, PCA is susceptible to yielding meaningless Euclidean distances between categories and not being able to capture the essence of the categorically dominated training data set. An improvement may be to reduce dimensionality through mutliple correspondence analysis (MCA) instead of PCA.