Technical University of Denmark

02450
Introduction to Machine Learning and Data Mining

Report 2

April 23, 2021

Contents

1 Introduction

The present report is a continuation of the first report [1], where the main principles of machine learning techniques were introduced and a selected data set was prepared to apply them. The Framingham Heart Study is chosen as data set, which is an ongoing cardiovascular study on residents of the town Framingham in Massachusetts [2].

This second report consists of two main parts, where a regression and a classification problem are solved for the Framingham data set. At first, the regression problem is solved by using three different models: a regularized linear regression model, an artificial neural network (ANN) and a baseline. Similarly, three methods are used for the classification as well, which are: a logistic regression, an artificial neural network and a baseline. The evaluation and comparison of the results is achieved by applying a K-fold cross-validation to obtain the corresponding errors in each model. Finally, a statistical evaluation takes place to estimate the performance of both the regression and classification models.

2 Regression

2.1 Regression: Part A

To begin with, the most elementary regression model, linear regression, is implemented for the selected data set. Initially, the attribute *heartRate* was selected for regression, as it is already mentioned in the first report [1]. However, every ratio attribute was tested with a linear regression model according to all the remaining attributes. As it can be seen in Fig. 1, *heartRate* distribution between real and predicted values is not ideal for this purpose. On the contrary, one can observe that the distributions of *sysBS* (top right) and *diaBP* (bottom left) are more suitable to implement regression, since the relation between real and predicted values fits significantly better along y = x axis. In that way, the systolic blood pressure attribute *sysBP* is selected to be predicted in our linear regression model.

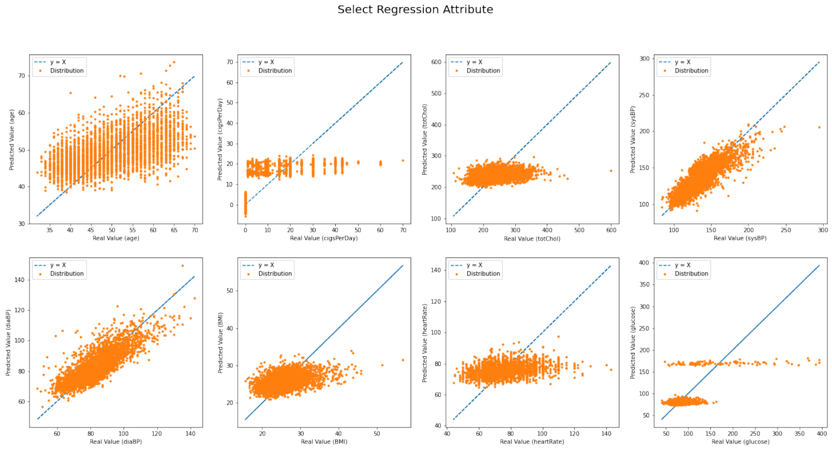

Figure 1: Attribute selection for regression.

Moreover, the data set should be transformed, so the regression is applicable. Firstly, one-of-K coding is applied to the education attribute which takes values from 1 to 4, representing lower to higher level of education. The systolic blood pressure attribute *sysBP* is removed from the data matrix **X**, since this is the variable that is going to be predicted. Finally, the data matrix **X** is transformed in a way that each column (attribute) has mean value and standard deviation equal to 0 and 1, respectively.

Subsequently, the regularization parameter λ is introduced and defined within the range of 10^{-5} to 10^{8}. For each value of λ, a K-fold cross-validation is implemented with K = 10, to estimate the generalization error. At first, a feature selection is done to select a subset of attributes and, in that way, reduce the model complexity. The K-fold cross-validation is initially applied with all the attributes considered and, as it can be observed in Fig. 2, several features could be disregarded such as *currentSmoker*, *cigsPerDay* etc.

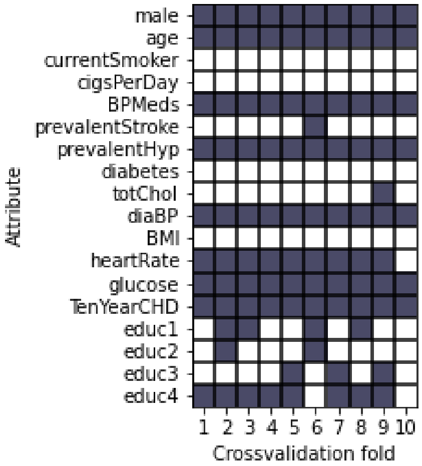

Figure 2: Feature selection.

In that way, the initial set of 18 attributes is reduced to a subset containing only 9 features. In Fig. 3 (left pane), the weights of each attribute of the subset are plotted for the different values of regularization factor λ. It can be observed that when λ is small, the weights take high values, showing high variance and low bias. On the other hand, as λ becomes larger the weights are dragged towards the x-axis and they appear to have low variance but high bias. Additionally, the generalization error is estimated for both training and test set in the aforementioned range of λ values and the result is shown in the right pane of Fig. 3. In general, we see that the training and test errors follow the same pattern as they remain steady for the lower values of λ and then significantly increase when λ becomes larger.

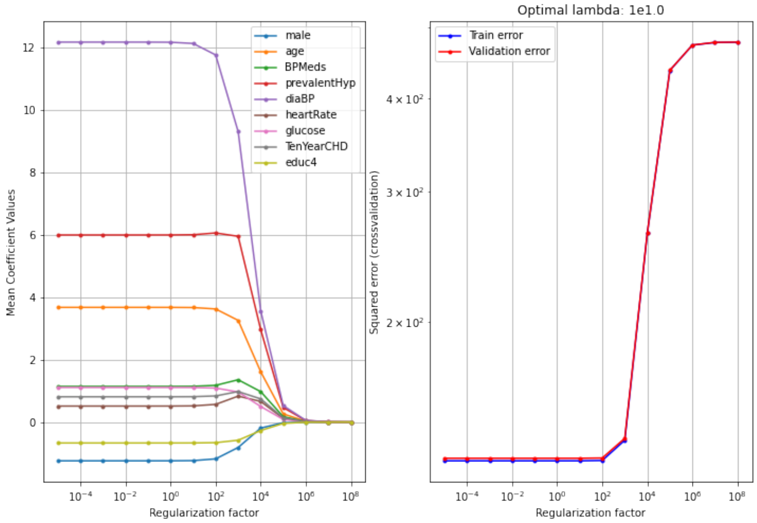

Figure 3: (a) Weights of the selected features as a function of λ; (b) MSE as a function of λ for training and test set.

In Fig. 4, the mean squared error (MSE) for both the train and the test set is calculated in each fold. It can be seen that for K = 2 the test error appears to have its lowest value, which is E = 119.41.

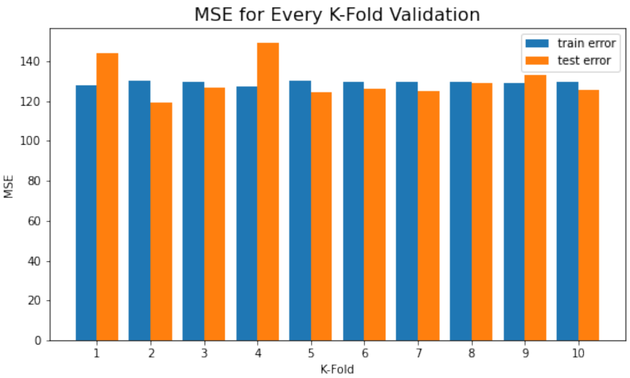

Figure 4: MSE for train and test sets in each fold.

The systolic blood pressure is predicted according to the equation below, with a set of weights calculated for each attribute:

$$\begin{aligned} sysBP = 132.21 &- 1.32 \cdot male + 3.64 \cdot age + 0.95 \cdot BPMeds + 5.77 \cdot prevalentHyp \\ &+ 12.54 \cdot diaBP + 0.49 \cdot heartRate + 1.14 \cdot glucose + 1.02 \cdot TenYearCHD \\ &- 0.71 \cdot educ4 \end{aligned}$$

Lastly, as it can be observed in Fig 5a, both true and predicted values have similar results, as the distribution follows the y = x line with a small variance as it seen in Fig 5b.

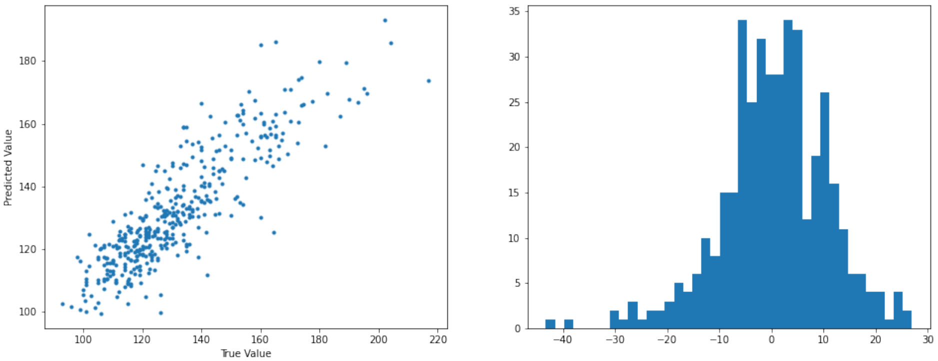

Figure 5: (a) True vs Predicted Value; (b) Residuals of distribution

2.2 Regression: Part B

2.2.1 Models Description

In this part, three different methods are implemented to the specific data set, in order to make a prediction to the systolic blood pressure attribute, as in the section above. An artificial neural network, a regularized linear regression model and a baseline model are the methods to be used. For the evaluation of the models, a two-level cross-validation technique is applied with $K_1 = K_2 = 10$. It is worth mentioning that all the models are trained and evaluated using the same training and test set.

For the artificial neural network a set of hidden nodes is used to find the optimal ones. With exception of the first and the second fold where the optimal nodes are ten, all the remaining folds used eight hidden nodes as the optimal choice. In addition, the maximum iterations are set at 10,000, where the outer cross validation fold does not need so many iterations to converge. During this trial and error phase, it is noted that as the hidden nodes are increasing, the generalization error follows the same increasing trend.

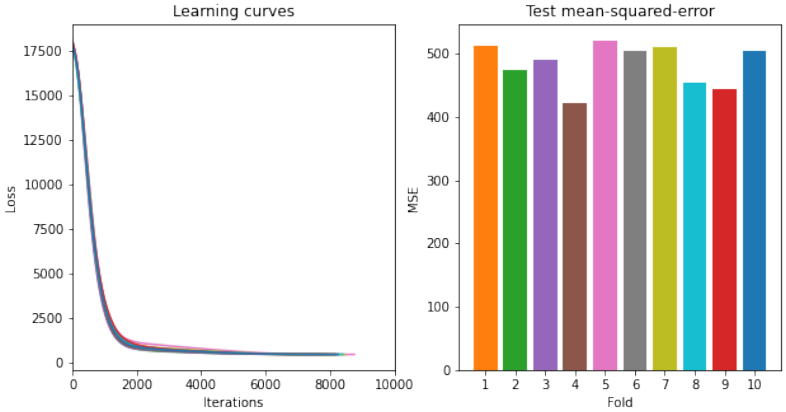

Figure 6: (a) Learning Curve for Every Outer Fold; (b) MSE for Every Outer Fold.

Regarding the regularized model, as also explained in the section above, a regularization parameter λ is introduced and tested within each cross validation for a range of values between $[10^{-5}, 10^{8}]$. The minimum generalization error is found for $\lambda = 10$.

Lastly, the baseline model is just a simple linear regression with no features. The way the model works is by computing the mean value of the systolic blood pressure in every training set in each fold. The baseline model is used to provide a minimum evaluation level for the other models.

2.2.2 Results Comparison

In the table below, the main results regarding the generalization error of the three models are summarized. For all the models, the generalization error is selected as a metric in order to compare them. The generalization error for a regression problem is the squared loss per observation.

$$E = \frac{1}{N^{test}} \sum_{i=1}^{N^{test}} (y_i - \hat{y}_i)^2 \tag{1}$$

The results show that the best performing model is the regularized linear regression model with the smallest generalization error equal to 130.51. The remaining models show almost the same results with a generalization error of 483.22 and 481.64 for the ANN and the baseline model, respectively.

Fold	**ANN**		**Regularized Linear Regression**		**Baseline**
i_{th}	h_i	E_i	λ_i	E_i	E_i
1	10	512.04	10	159.39	550.36
2	10	473.35	10	119.29	466.74
3	8	489.48	10	128.21	488.50
4	8	420.96	10	110.27	417.54
5	8	520.01	10	135.22	510.33
6	8	504.71	10	130.73	499.95
7	8	510.67	10	155.53	500.82
8	8	453.95	10	134.17	448.64
9	8	443.11	10	112.18	435.00
10	8	503.94	10	120.13	498.60

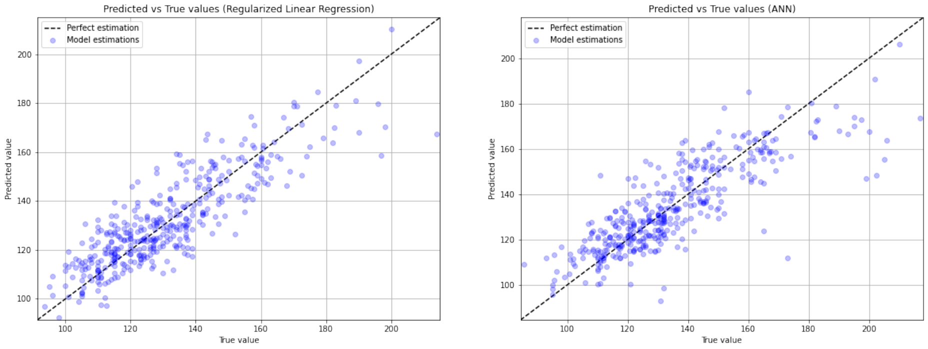

Figure 7: (a) Predicted vs True Value for the Linear Regression Model; (b) Predicted vs True Value for the Artificial Neural Network.

2.2.3 Statistical Evaluation

Finally, a paired t-test is applied to statistically compare the performance of the three models using the setup I. As it can be seen in the table below, the linear regression model show small p-values when it is paired with the other two models. Thus, it can be concluded that the linear regression model has significant performance difference with the ANN and the baseline models. On the other hand, the pair of ANN with the baseline model show larger p-value, meaning that they have more similar performance, but since the confidence interval does not include zero, these models cannot have the same performance.

Pairwise Models	Confidence Interval	P-Value
ANN/Regularized Linear Regression	[304.061, 463.811]	1.42e-19
ANN/Baseline	[-9.859, -1.083]	0.0073
Baseline/Regularized Linear Regression	[300.724, 456.206]	5.51e-20

3 Classification

In this part, a classification problem is solved for the selected data set and the results are statistically evaluated. Similar to the previous section, three methods are used: a baseline, a logistic regression, and an artificial neural network (ANN). A two-level cross-validation is implemented to compare the performance of the aforementioned models. In this case, the classification problem is binary, since it predicts whether or not a patient will develop a coronary heart disease in the next ten years.

3.1 Models Description

For logistic regression, the regularization parameter λ is used once again as a complexity-controlling parameter and its values are defined within the range $[10^{-8}, 10^{5}]$. Regarding the ANN, a similar to the regression problem procedure is applied, hence the selected complexity parameter is again the hidden units h. Several values are examined as h takes values equal to 1, 5, 10 and 15. In Fig. 8, the optimal weights of each attribute are shown as well as the optimal value of hidden units, which is obviously h = 1.

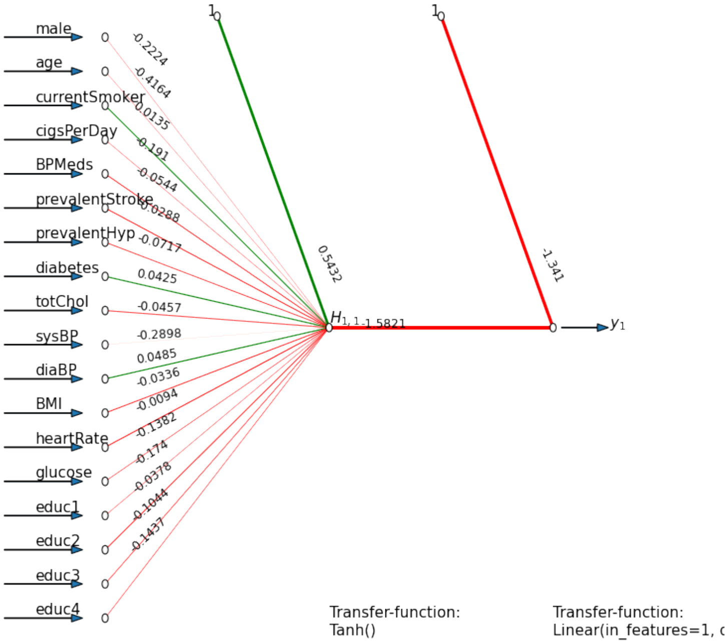

Figure 8: Neural net class.

In addition, the learning curves (left pane) as well as the mean-squared-error (MSE) in each fold (right pane) are plotted in Fig. 9. The learning curves show the loss in respect with the iterations, where the maximum iterations are set to 10,000. Regarding the MSE, its lowest value is observed in K = 4 fold.

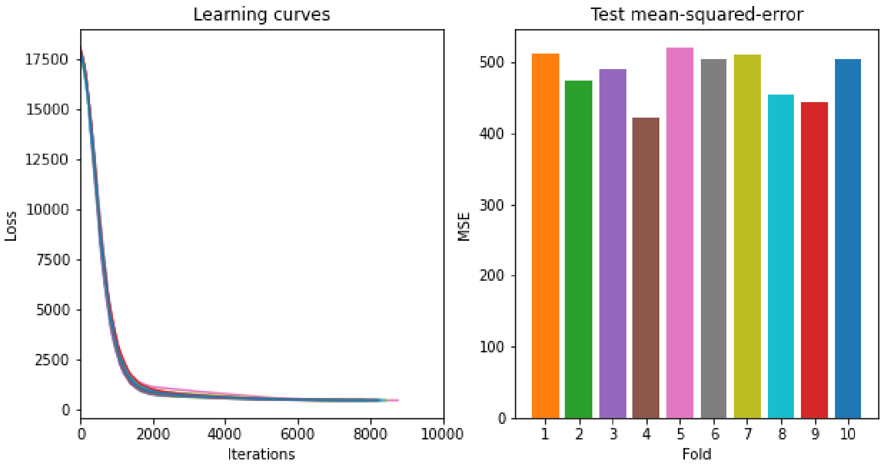

Figure 9: (a) Classification error for logistic regression; (b) Classification error for ANN.

Finally, the baseline model for classification follows the same purpose as in the regression model. It is a model with no features which computes the largest class of *TenYearCHD* (0 or 1) on the training data, and then it predicts all the values that belong to this specific class in the test data.

3.2 Results Comparison

As already mentioned, a two-level cross validation is applied in order to compare the three models. This is achieved by comparing the error rate, which in this case, is given by the expression:

$$E = \frac{\{\text{Number of misclassified observations}\}}{N^{test}} \tag{2}$$

The selected parameters and the results of the two-level cross-validation for each of method are summarized in the table below in each fold. In addition, the error rate is plotted in Fig. 10 for both the logistic regression and the artificial neural network (ANN). It can be observed that both models appear to have the lowest test error for K = 5 with value E = 0.126 in both cases. This is also the case for the baseline model, where the error rate is E = 0.121.

Fold	**ANN**		**Regularized Logistic Regression**		**Baseline**
i_{th}	h_i	E_i	λ_i	E_i	E_i
1	5	0.163	10	0.146	0.171
2	1	0.139	10	0.141	0.144
3	1	0.173	10	0.168	0.168
4	1	0.141	10	0.141	0.149
5	1	0.126	10	0.126	0.121
6	1	0.139	10	0.136	0.146
7	1	0.183	1e−5	0.178	0.183
8	1	0.131	1e−5	0.134	0.139
9	1	0.127	10	0.129	0.136
10	1	0.127	100	0.129	0.141

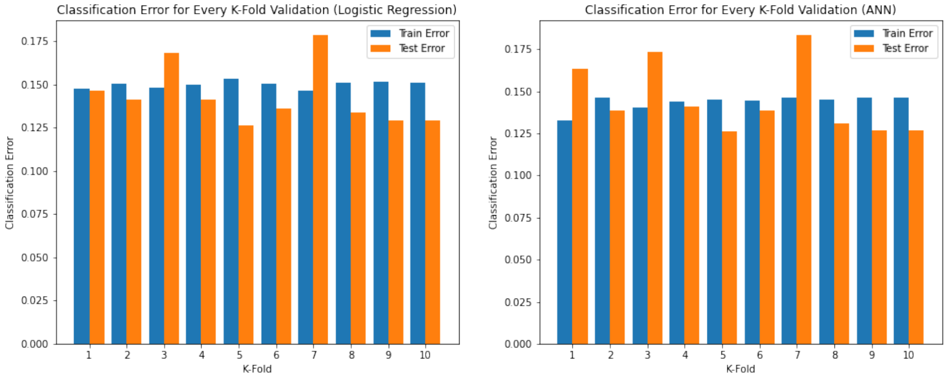

Figure 10: (a) Classification error for logistic regression; (b) Classification error for ANN.

Finally, the classification problem that predicts whether a patient will develop a coronary heart disease in the next ten years as well as the corresponding weights calculated above can be expressed using the equation:

$$\begin{aligned} TenYearCHD = &-1.997 + 0.23 \cdot male + 0.52 \cdot age + 0.022 \cdot currentSmoker \\ &+ 0.236 \cdot cigsPerDay + 0.039 \cdot BPMeds + 0.068 \cdot prevalentStroke \\ &+ 0.105 \cdot prevalentHyp + 0.014 \cdot diabetes + 0.069 \cdot totChol \\ &+ 0.305 \cdot sysBP - 0.023 \cdot diaBP + 0.013 \cdot BMI - 0.018 \cdot heartRate \\ &+ 0.155 \cdot glucose + 0.042 \cdot educ1 - 0.063 \cdot educ2 - 0.004 \cdot educ3 \\ &+ 0.031 \cdot educ4 \end{aligned}$$

3.3 Statistical Evaluation

At last, a statistical evaluation is applied once again, similar to the regression part. Also in this case, the selected test to evaluate the performance of the models is the setup I with $\alpha = 0.05$. According to the results of the table below, there is a strong chance that both the ANN and the baseline model will have the same performance due to the fact that the confidence interval contains the value 0 and the p-value is quite large. However, this is not the case for the other pairs, since the p-values are significantly smaller.

Pairwise Models	Confidence Interval	P-Value
ANN/Regularized Logistic Regression	[0.003, 0.027]	0.03125
ANN/Baseline	[-0.002, 0.007]	1.0
Baseline/Regularized Logistic Regression	[-0.025, 0.000]	0.125

4 Discussion

The main focus of this project was to implement two different machine learning methods, regression and classification for a selected data set. Regarding regression, the scope of the analysis was to make a sufficient prediction to the systolic blood pressure. The classification method aimed to predict whether patients will suffer from a coronary heart disease in the upcoming ten years, as it was intended for the data set in the first place.

Considering the regression problem, the model will only show better results with a different data collection that will aim in different attributes, since the current ones are not strong correlated.

On the other hand, the classification problem yielded some positive and reliable results, by creating a high-accuracy model that aimed to predict the patients with a higher probability of developing a coronary heart disease.

Lastly, the two-layer cross-validation that was applied to accurately estimate the generalization error was of major importance. It made the predictions more reliable when new data would be given as inputs, where it is deemed necessary, especially when the predictions relate to medical issues.

Comparing the results of this project with other studies [3], we can conclude that not only did we achieved similar or better results, but also made a similar analysis in regard to the correlation and the weights of the attributes.

References

[1] C. Papadopoulos, E. Lydakis Simantiris. *Introduction to Machine Learning and Data Mining - Report 1*, Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2021.

[2] Kaggle, Framingham Heart study dataset, viewed 10 February 2021, https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset.

[3] Research Gate, Healthcare Analytics: A Case Study Approach Using the Framingham Heart Study, viewed 15 April 2021, https://www.researchgate.net/publication/336703415-Healthcare-Analytics-A-Case-Study-Approach-Using-the-Framingham-Heart-Study.