Technical University of Denmark

Machine Learning and Datamining

Cleveland database : Regression and Classification

Authors:
Victor Niaussat

April 20, 2021

Contents

1 Introduction

In report 1 initial visualization, including principal component analysis (PCA) was applied to the Cleveland database of clinical results of 303 patients undergoing angiography at the Cleveland Clinic (Cleveland, Ohio) between May 1981 and September 1984.

In this report we will apply various machine learning methods for regression and classification to the dataset and evaluate their performance using 2-level cross validation and statistical tests.

In agreement with previous studies performing classification tasks on the dataset to identify patients with and without heart diseases, it was observed in report 1 that the two groups could with good approximation be separated in the PCA. In this report we aim to apply various machine learning models to the dataset in order to classify patients with or without heart disease using logistic regression and artificial neural networks.

To the best of our knowledge no previous studies using regression models to the dataset have been published. In report 1 very weak correlation between the different attributes was observed, however it was found, that the maximum hear rate seemed to be somewhat correlated with some of the other attributes, in particular age. In this report we will attempt to use regularized linear regression and artificial neural networks to predict this value based on the remaining attributes in the dataset.

2 Regression part A

The purpose of this linear regression is to estimate the maximum heart rate of the patient based on the other attributes in the data set. As a rule of thumb, the maximum heart rate of a normal, healthy person can be calculated from the formula

$$\text{maximum heart rate} = 220 - \text{age}$$

suggesting that there is a linear relation between the age and the maximum heart rate[1]. It would be interesting to see how this relation fares when the person is not necessarily healthy, but for example has increased cholesterol or the presence of heart disease.

The attributes *Chest Pain*, *Thalassemia* and *Slope* are nominal and contain multiple categories, and we therefore apply a one-out-of-k encoding to these features. Furthermore, the data matrix is standardized such that each column has a mean of zero and standard deviation of one, while the feature to predict remains unchanged. Finally, a linear regression model is fitted on the data, and the estimated results from the regression model is matched with the true values of the maximum heart rate in figure 1. Note this regression model has not been post processed or tested using crossvalidation, and is likely to be overfit to the data.

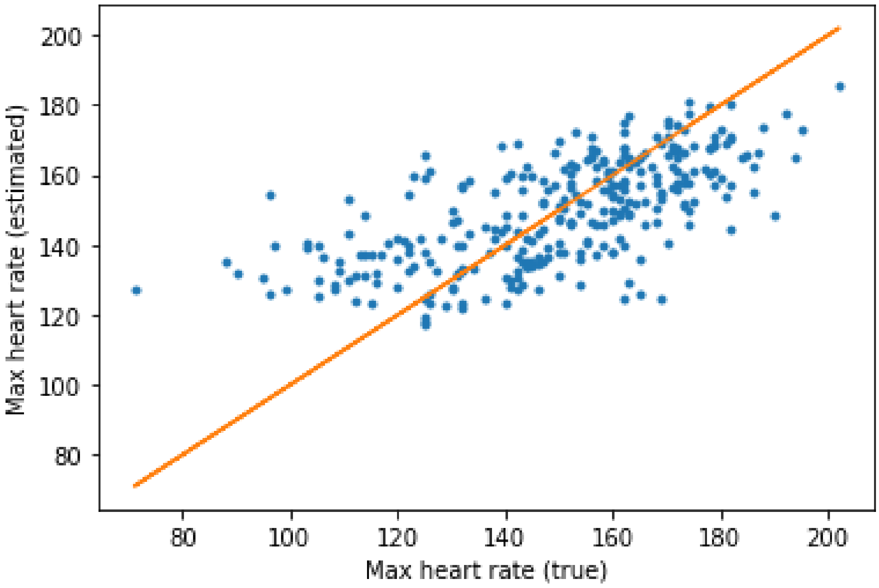

Figure 1: Linear regression estimations and real values

while a rough linear trend can be seen on the plot, the data points still deviate a lot from the orange line, which indicates the output for a perfect regression model. Regardless, to make this regression model better for any new inputs of data, regularization can be applied. Regularization is performed by introducing a regularization parameter λ to the weights, and then use cross-validation to calculate the validation error of trained models on the test sets for different values of λ. This is done in the script `regression_part_a_2.py`, and a plot of the regularization of the linear regression model is shown in figure 2.

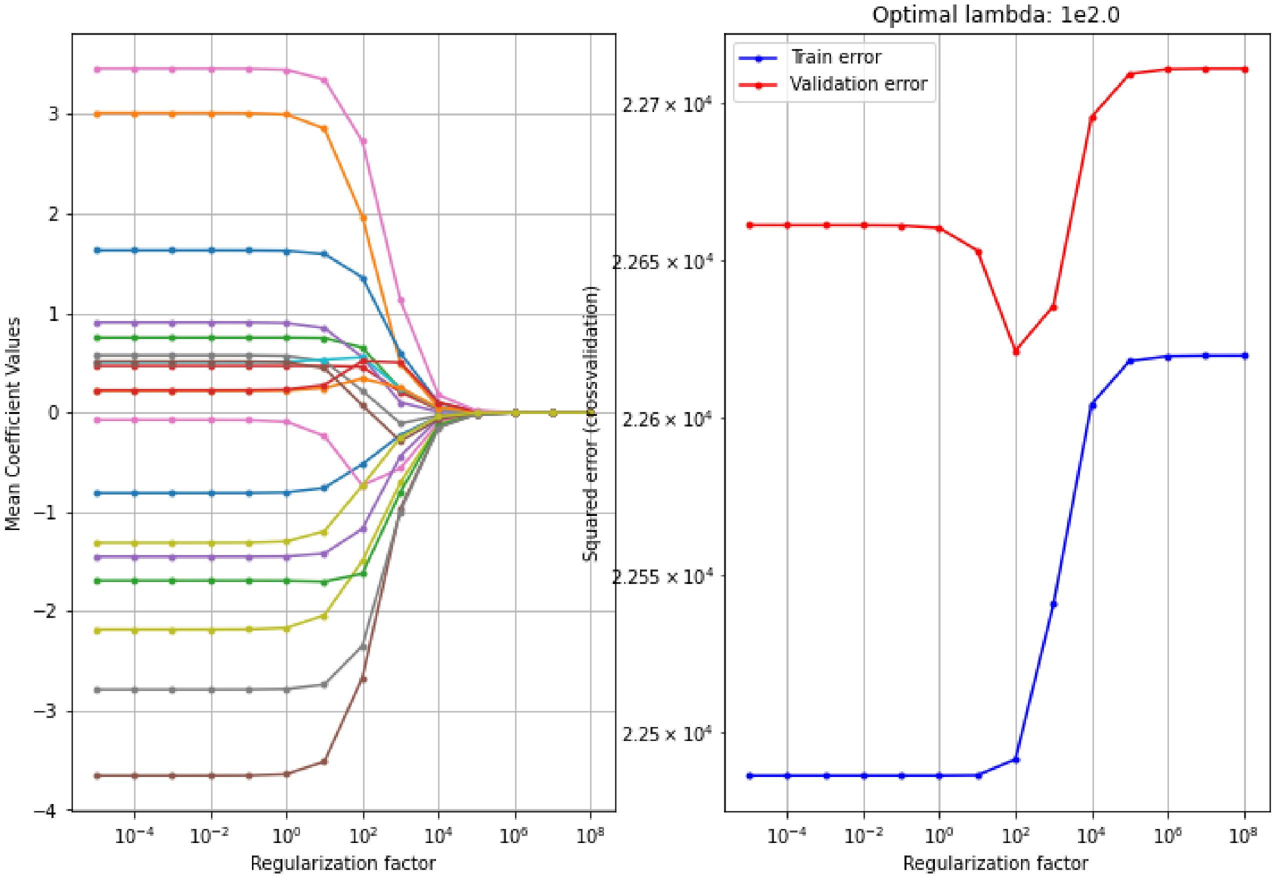

Figure 2: Linear regression regularization results

In this regularization, the λ values are powers of ten in the range $[10^{-5}; 10^{8}]$. The validation error is the red line on the right plot, which shows the sum-of-squares error when the regression models are applied to test data. The lowest validation error is found when $\lambda = 10^2$. This means that the coefficients of the weights should actually be reduced to provide the best result when the model is applied onto new data. This results in bias increasing and variance reducing, indicating that the original model was overfitted to the data. This is shown in the plot to the left in figure 2.

The regularization also shows, that the best model from the cross validation is one where the weights of each observation are as in table 1.

Feature	Weight
Age	-10.76
Sex	-0.54
Resting_Blood_Pressure	2.04
Colestrol	0.68
Fasting_Blood_Sugar	0.46
Rest_ECG	0.58
Exercised_Induced_Angina	-2.73
ST_Depression	-0.72
Major_Vessels	0.24
Target	-1.54
Chest_Pain_1	0.58
Chest_Pain_2	1.35
Chest_Pain_3	0.34
Chest_Pain_4	-1.63
Thalessemia_3	0.5
Thalessemia_6	-1.2
Thalessemia_7	0.1
Slope_1	2.78
Slope_2	-2.4
Slope_3	-0.74

Table 1: Weights of observations in linear regression model

The maximum heart rate of each new data point will be calculated as the sum of the weight of the attribute times the value. Here it can be seen that the attribute with the most weight is the age of the patient, and that the relation is negative. This is in agreement with the general rule of thumb, predicting that the maximum heart rate can be calculated as a function of age alone. The other attributes do not contribute significantly to the result.

3 Regression part B

In this section, we will compare the regularized linear regression model from the previous section with two additional models; a neural network (NN) and a baseline. Two level cross validation and paired t-tests will be used to compare the models.

In interest of time, 5-fold cross validation is used to split the data in a training set and a test set. The models are then trained on the training set and evaluated on the test set 5 times using different values of complexity controlling parameters (the regularization parameter, λ, for the linear regression and the number of hidden units,n for the neural networks), using the same data split for all the three models, which will be explained in the following.

The baseline model is a linear regression model with no features, i.e. a model that calculates simply the mean of the data.

The linear regression model (as explained in the previous section) is trained for 10 values of regularization parameters, λ, rangin from 10^{-1} to 10^{10}

For the neural network `the train_neural_net` function from `toolbox02450` is used, however the code is modified to select the model with the lowest test error rather than training error to avoid over training. Based on a number of test runs, a model with three hidden layers with $n, 2n$ and $4n$ as the number of hidden units respectively. As the complexity controlling parameter the number of hidden units is used. Based on the test runs it is found that for large values, the model is rapidly over trained giving very high test losses, whereas for very low numbers of hidden units the network simply predicts a value close to the mean, similarly to the baseline model. In the 2 level cross validation loop, the values $n = 1, 2, 10, 20$ and 50 are selected for evaluation.

The best parameters in each fold along with the corresponding test mean square error (MSE), E^{test}, are given in Table 2

Outer fold	ANN		Linear regression		Baseline
i	E_i^{test}	n_h	E_i^{test}	λ^*	E_i^{test}
1	560.9	2	398.3	1000	560.6
2	465.5	2	400.0	100	458.8
3	485.9	2	396.5	100	591.5
4	481.8	50	391.4	10	385.3
5	407.9	2	392.4	1000	388.8

Table 2: 2 level 5-Fold cross-validation table used to compare the three regression models

From the table, it is clear that the MSEs of all the tested regression models are large. The linear regression model has slightly lower error rates on average than the baseline and the neural network. The best values for lambda range between 10^1 to 10^3. The neural network has MSEs very similar to those of the baseline, which is simply predicting the mean of the data. The majority of the trained models find $n = 2$ as the best number of hidden units, however the MSEs of the different values in each fold are similar, regardless of the selected number of hidden units. For illustration, figure 3 shows the learning curves and losses of neural networks ($n = 2$) trained for 3 different training sets generated by 3-fold cross validation. Note that these are not the learning curves used in the 2-level cross validation. Figure 4 shows the predicted values against the true values for the last fold of the demonstration. It is clearly observed that the neural network simply learns to predict a value close to the mean of the data, which explains the similarity with the baseline loss.

To statistically compare the performance of the three models (Table 2), the paired t-test is applied to the neural network against the baseline, the regularized linear regression against the baseline and the neural network against the linear regression. In the t-test it is determined whether the difference in the mean of the predicted data is statistically significant. The output of the test is the p-value, which is a measure of the unlikeliness of a given observations given that the null hypothesis is true, that is that the two tested models are identical. Hence, the smaller p-value the more likely is it that the null-hypothesis (that the difference in mean is 0) is wrong. The tests are here performed at a significance level of $\alpha = 0.05$, meaning that the null hypothesis is rejected if the p-value fall below this value. The resulting p-values are given along with the corresponding confidence interval, CI, in Table 5 for the last fold of Table 2.

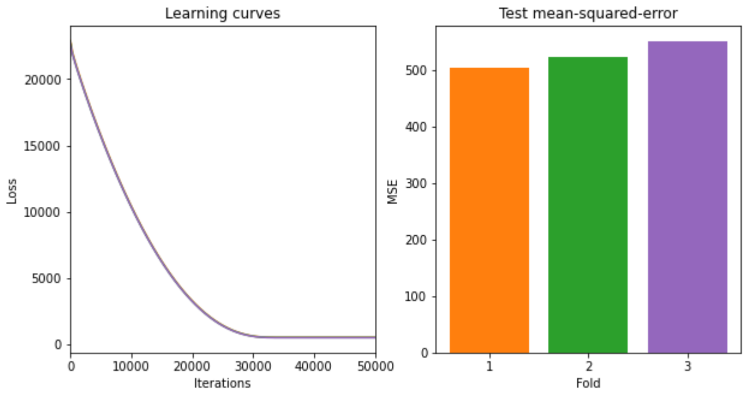

Figure 3: Learning Curves and test loss for 3 neural networks using $n = 2$ and 3-fold cross validation

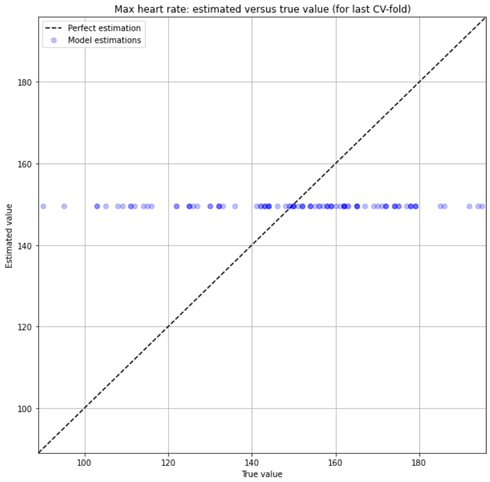

Figure 4: predicted max heart rate against the true value for the last model in Figure 3

	p	CI
NN vs Lin	0.14	(-413;123)
NN vs Base	0.014	(19; 344)
Lin vs Base	$7.6 \cdot 10^{-5}$	(165;488)

Table 3: paired t-test, comparing the three regression models of the last fold

It is found, that only the first null hypothesis (comparing the linear regression and the neural network) is not rejected at the given significance level. Based on this statistical test, the regularized linear regression model should be selected as the best performing model for the regression problem, however as the error rates for all models are high, none of the models seem to be well suited for the problem.

4 Classification

In this part, a comparison is made between 3 different classification models. The purpose of the classification in this part is to separate the data-set in two parts: patients which have a heart disease and patients which do not have a heart disease. It corresponds to the attribute *target* in the data-set *Heart Disease Data Set*. Hence, it is a binary classification problem because the attribute *target* is a binary variable.

We will compare logistic regression, artificial neural networks for classification and a baseline.The baseline is a model which compute the largest class on the training data, and predict everything in the test-data as belonging to that class (corresponding to the optimal prediction by a logistic regression model with a bias term and no features). For logistic regression, we will once more use λ as a complexity-controlling parameter. For the artificial neural network for classification model, the relevant complexity controlling parameter is the number of hidden units n_h and the range of values we selected is $n_h \in [\![1;10]\!]$

Then, it trains classification models and uses a 10-fold two-level cross validation on each classification model (Artificial Neural Networks ANN, Logistic regression and Baseline). The figures below show the best learning curves for a 10-fold two-level cross validation of the ANNs with 1 hidden layer and 10 hidden units:

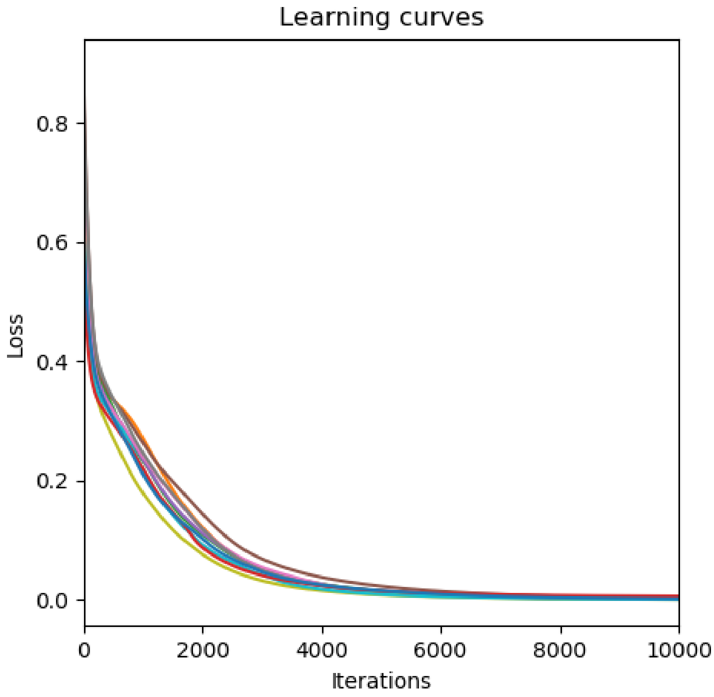

Figure 5: Best learning Curves for each fold for a 10-fold two-level cross validation on a ANN

The table below summarises each cross validation fold, the values of n_h and λ^* (the value of λ which minimizes the error rate on logistic regression model for classification) and errors test, done in the script`project2_Classification.py`:

Outer fold	Neural Network		Logistic Regression		Baseline
i	E_i^{test}	n_h	E_i^{test}	λ^*	E_i^{test}
1	0.418	1	0.136	1,00e+3	0.456
2	0.398	2	0.151	1,50e+2	0.448
3	0.288	3	0.138	3,40e+1	0.451
4	0.288	4	0.142	1,50e+2	0.435
5	0.286	5	0.127	1,30e+2	0.446
6	0.218	6	0.148	2,50e+1	0.467
7	0.232	7	0.13	1,00e+3	0.462
8	0.207	8	0.135	1,30e+2	0.436
9	0.165	9	0.164	1,30e+2	0.451
10	0.225	10	0.162	1,90e+1	0.455

Table 4: 10-Fold cross-validation table used to compare the three models in the classification problem.

From the table we gather that the artificial neural network is better when it increases the number of hidden units, since the test error decreases. Moreover, it seems that Logistic regression performs better than the artificial neural network and the baseline.

Then, our goal is to compare our classifiers. To do that, we perform a statistical evaluation of our three models and compare the three models pairwise. Specifically, we want to know if a model A is better than a model B. The McNemar's test compares two models giving $\hat{\theta}$, a confidence interval and a p-value. Let θ_A denote the (true) chance classier A is correct and θ_A the (true) chance classier B is correct. The difference in performance is given by

$$\hat{\theta} = \theta_A - \theta_B$$

.

If $\hat{\theta} \geq 0$, model A is preferable over B. Moreover, the lower p is, the more evidence there is A is better than B, but we only interpret the p-value together with the estimate $\hat{\theta}$ and ideally the confidence interval computed above. The table below summarizes results from McNemer's test with significance level of $\alpha = 0.05$.

Model B Model A	Neural network (n_h=10)			Logistic Regression			Baseline		
	$\hat{\theta}$	p	CI	$\hat{\theta}$	p	CI	$\hat{\theta}$	p	CI
Neural Network	-	-	-	-0.14	9.62e-08	[-0.191; -0.091]	0.14	1.11e-06	[0.084; 0.191]
Logistic regression	0.14	9.62e-08	[0.091;0.191]	-	-	-	0.279	1.69e-13	[0.209;0.347]
Baseline	-0.14	1.11e-06	[-0.191;-0.084]	-0.279	1.69e-13	[-0.347;-0.209]	-	-	-

Table 5: $\hat{\theta}$, confidence interval and p-value from McNemera's test

This test ensures that Logistic regression is better than the artificial neural network, better than the baseline. p-values are very low and each bounds of each confidence interval ensures the validity and the good interpretation of $\hat{\theta}$ value.

In the second row of the table, the logistic regression is compared with the other two models.It can be seen that the difference in performance between the logistic regression and the other models is positive for both tests. Hence, the Logistic regression is better than the other. The two models are better than the baseline. Hence, to classify *Heart Disease Data Set*, it preferables to use the Logistic regression to separate the data-set in two parts: patients which have a heart disease and patients which have not heart disease.

5 Discussion

In this report we have applied various regression and classification models to the Cleveland database. As expected from the initial visualization in report 1 as well as the results from previous publications, the classification models succesfully classify patients with and without heart diseases, with misclassification rates around 14 % for logistic regression and 25 % for the neural networks which is significantly better than the baseline model, which has a misclassification error around 45 %. Moreover, McNemar's test ensures that Logistic regression is better than the artificial neural network, better than the baseline with a high degree of certainty thanks to low p-values .

In a previous study by Detrano et. al [2], logistic regression was used to classify the data. In this study the model was evaluated on new datasets, rather than test-sets generated by K-fold cross validation. depending on the disease prevalence in the test group and the decision threshold the obtained misclassification rates range between 18 and 42 %, hence the misclassification rate obtained in this report compare well with the results of this study.

Interestingly the neural network does not perform significantly better than the logistic regression. As the neural network employed herin is very and migh e significantly improved, e.g. by adding additional hidden layers.

In the regression part both artificial neural networks and regularized linear regression were found to perform poorly. The Linear regression model was found to yield the lowest mean square errors of the three, however in a statistical test, it is found that the difference between the neural network and the linear model is not siginificant. The neural network yields results similar to those of the baseline model predicting the mean of the data. The poor results are in agreement with expectations, considering that the dataset has previously only been used for classification in the literature and that only a very weak correlation between the attributes was observed in report 1.

References

[1] Centers for disease control and prevention. *Target Heart Rate and Estimated Maximum Heart Rate.* URL: https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.html.

[2] Robert Detrano et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease". In: *The American Journal of Cardiology* 64.5 (1989), pp. 304–310. ISSN: 00029149. DOI: 10.1016/0002-9149(89)90524-9.